Freeform Search

Database: US Pre-Grant Publication Full-Text Database
US Patents Full-Text Database
US OCR Full-Text Database
EPO Abstracts Database
JPO Abstracts Database
Derwent World Patents Index
IBM Technical Disclosure Bulletins

Term: L8 and phrase$

Display: 50 **Documents in Display Format:** - **Starting with Number** 1

Generate: ○ **Hit List** ◉ **Hit Count** ○ **Side by Side** ○ **Image**

Search Clear Interrupt

Search History

DATE: Monday, May 23, 2005 Printable Copy Create Case

Set Name side by side	Query	Hit Count	Set Name result set
DB=PGPB,USPT,USOC,EPAB,JPAB,DWPI,TDBD; PLUR=YES; OP=OR			
L9	L8 and phrase$	29	L9
L8	L7 and (sentence near extract$)	32	L8
L7	L1 and (related near document)	240	L7
L6	l1 and (temporal near process$)	0	L6
L5	L4 and root$	0	L5
L4	L3 and phrase$	8	L4
L3	L2 and (sentence near extraction)	14	L3
L2	L1 and (collection near document$)	220	L2
L1	summary near document$	1058	L1

END OF SEARCH HISTORY

L9: Entry 22 of 29 File: USPT Jun 17, 2003

DOCUMENT-IDENTIFIER: US 6581057 B1
TITLE: Method and apparatus for rapidly producing document summaries and document browsing aids

Brief Summary Text (3):
This invention relates to computer-assisted information storage and retrieval and, more particularly, to producing document summaries and document browsing aids.

Brief Summary Text (5):
As part of search results corresponding to a user query, for example, in an information retrieval system, a query-biased summary generation system provides a document summary that incorporates sentences, sentence fragments, or text spans that are relevant to the user query. The full text of the document must be available in order to create the query-biased summary. Usually, the summary includes the sentences having the greatest number of user query terms that appear most frequently. The summary can also include sentences that are closely related to the query by incorporating synonyms of the query terms into the criteria for the selection of the included sentences. With the current state of the art, the generation of a query-biased summary requires significant processing time.

Brief Summary Text (6):
Current information retrieval systems and information management systems, such as web catalogs, search engines, and document indexes, do not use query-biased summaries. Topical document summaries that are relevant to the user query are not provided. Instead, for example, they present the first few sentences of a document as an indication of the content of that document. These first few sentences may be extracted from the document and stored as a summary of that document for later use in response to a user query. While this technique works well with news stories that use the inverted pyramid style of writing, where the most important facts are mentioned toward the beginning of the article, it does not work well with other text genres that typically do not use the inverted pyramid style.

Brief Summary Text (10):
Current web search engines do not provide query-biased document summaries for several reasons. The main reason is that computation time is extremely limited. Since revenue generation is dependent on advertising exposures, response time to a query is critical. Generating query-biased summaries as part of the retrieval process would add enough of a delay to decrease the revenue throughput of the web catalog. The added delay might additionally cause some users to switch to a competitor's faster web catalog. Moreover, web catalogs answer tens of millions of queries per day, and adding a second or two of computation time per query might necessitate the purchase of additional equipment to handle the increased demands on the system.

Brief Summary Text (14):
It is an object of the present invention to provide a method and apparatus for rapidly producing document summaries and document browsing aids which do not require storing the full text of the documents.

Brief Summary Text (15):
Accordingly, we have developed a method and apparatus for rapidly producing document summaries and document browsing aids by, at index creation time, precomputing and caching query relevant information required for creating the summaries.

Brief Summary Text (17):
In the specification and claims, the word "term" means single words, word n-grams, and/or phrases. An "n-gram" is a string of characters that may comprise all or part of a word.

Brief Summary Text (22):
The present invention can be adapted to any text summarization method involving text spans and to a variety of computation intensive document browsing aids. The invention significantly reduces the time it takes to produce a query-biased summary without substantially affecting the quality of the summary. Thus, the invention makes it feasible for web search engines and other information retrieval systems to display document summaries and other document browsing aids with the results of a search.

Brief Summary Text (24):
The present invention also applies to document browsing aids, such as keyword gists, thumbnail images, clustering, and categories. A keyword gist is a shortened form of a document in which all but the keywords have been deleted. A thumbnail is a reduced image of the document (e.g., a photo reduction). Clustering involves grouping related documents together into a cluster. Categorizing is similar to clustering, but instead assigns a label to each document in which the label identifies which group that document belongs. By optimizing the search time generation of these aids through the precomputing and caching of information, the present invention makes these aids practical for real world applications, such as web catalogs and document indexes.

Detailed Description Text (2):
Referring to FIG. 1, a first embodiment of a method according to the present invention begins at index creation time with step 110 by extracting information from a document 100. The extracted information is information that is relevant to at least one dummy query and is necessary to compile at least one temporary summary for the document 100. In step 112, information 113 is cached for later use in creating a summary of the document 100. At search time, a user query is entered at step 114, and step 116 generates the summary for the document 100 by utilizing the cached information 113.

Detailed Description Text (5):
Whole sentences may be extracted. In order to aid in the decision of which sentences to extract, a score may be assigned to each sentence of the document 100. For example, step 110 may extract at least the highest scoring sentence, or may extract a pre-defined number of the highest scoring sentences. Step 110 may extract the highest scoring sentences that have a score greater than a threshold score. Up to a pre-defined number of the highest scoring sentences having a score greater than a threshold may be extracted. Likewise, step 110 may extract all of the sentences of a paragraph that contain a number of the highest scoring sentences. Furthermore, each score may be assigned based upon the similarity of the sentence to sentences of documents in a results document list created by the dummy query.

Detailed Description Text (8):
A second embodiment of a method of the present invention is shown in FIG. 2. At index creation time, information from a document 200 which is relevant to at least one dummy query and is necessary to compile at least one temporary summary is extracted in step 210. Step 212 caches information 213. Step 214 compiles a summary of the document 200. Summary 217 is cached in step 216. At search time, a user query is entered at step 218. The cached summary 217 is utilized as the generated summary in step 220.

Detailed Description Text (9):
In a third embodiment of a method of the present invention, shown in FIG. 3, information from a document 300 that is relevant to at least one dummy query and is necessary to compile at least one temporary summary is extracted in step 310. Step 312 caches information 313. Step 314 generates links from the information 313 to the corresponding locations in the document 300 at which the information 313 is found. Step 316 caches links 317 that were generated in step 314. At search time, a user query is entered at step 318 and a summary of the document 300 is generated in step 318 using the information 313 cached in step 312. The links 317 may be provided with the summary to the user such that the user can jump directly to the relevant portions of the document 300.

Detailed Description Text (10):
FIG. 4 illustrates a fourth embodiment of a method of the present invention. At index creation time, at least two dummy queries are entered in steps 410a and 410b for use in steps 412a and 412b to extract information from a document 400 that is relevant to at least one dummy query and is necessary to compile at least one temporary summary. Steps 414a and 414b separately

cache the extracted information 415a and 415b, respectively. At search time, a user query is entered in step 416. Step 418 determines which dummy query best matches the user query. The dummy query that best matches the user query, as determined in step 418, determines whether step 420a or 420b generates the summary of the document 400 using the information 415a or 415b, respectively, cached in step 414a or 414b, respectively.

Detailed Description Text (13):
FIG. 5 shows a fifth embodiment of the present invention wherein dummy queries are utilized as in FIG. 4. At index creation time, dummy queries are entered in steps 510a and 510b for use in steps 512a and 512b to extract information from a document 500 that is relevant to at least one dummy query and necessary for compiling at least one temporary summary. Steps 514a and 514b separately cache extracted information 515a and 515b. However, once the information 515a and 515b is cached in steps 514a and 514b, steps 516a and 516b generate links from the query terms to the locations in the document 500 at which the query terms are found. Steps 518a and 518b cache links 519a and 519b generated in steps 516a and 516b, respectively. At search time, a user query is entered in step 520. Step 522 determines which dummy query best matches the user query. The dummy query that best matches the user query determines whether step 524a or 524b generates the summary of the document 500 using the information 515a or 515b, respectively, cached in step 514a or 514b, respectively. The links 519a and 519b may be provided to the user such that the user can jump directly to the relevant portions of the document 500.

Detailed Description Text (18):
In use, a document is stored on the server system 818. The document may be stored at a location other than the server system 818 as long as the server system 818 has the ability to access and retrieve information from the document. At index time, a method according to the present invention executes on the server system 818 to precompute and cache the information that is relevant to at least one dummy query and is required for creating a summary of or a document browsing aid for the document at search time. At search time, a user enters a user query on the client system 810. The communications links 812 and 814 relay this query across the network 816 to the server system 818. The method then generates the summary or document browsing aid in response to the user query and sends the results back to the client system 810 over the network 816 using the communications links 812 and 814.

Detailed Description Text (20):
The present invention permits document summaries and browsing aids to be rapidly produced by precomputing and caching query relevant information from a document at index creation time such that a summary or document browsing aid can be created at search time in a time-efficient manner from the cached information. By using the cached query relevant information at search time, the present invention also eliminates the need to have access to the entire document at search time in order to produce the summary or browsing aid.

CLAIMS:

1. A computer-assisted method for generating a summary of a document, comprising the steps of: at an index creation time, with access to the entire document extracting from the document information that is relevant to at least one dummy query and is necessary to compile at least one temporary summary, and caching at least part of the information comprising substantially less than the entire document; and at a later search time, generating the summary from the information cached.

14. The computer-assisted method according to claim 13, wherein: the information is extracted by the step of assigning at least one score to each sentence in the document according to the relevance of the sentence to the at least one dummy query, wherein at least the highest scoring sentence is extracted.

15. The computer-assisted method according to claim 14, wherein a pre-defined number of the highest scoring sentences are extracted.

17. The computer-assisted method according to claim 16, wherein up to a pre-defined number of the highest scoring sentences is extracted.

18. The computer-assisted method according to claim 14, wherein: the information is extracted by the steps of: assigning at least one score to each sentence in the document according to the

relevance of the sentence to the at least one dummy query, and extracting all of the sentences of a paragraph of the document which contains a number of the highest scoring sentences, and the summary generated is the paragraph.

27. The computer-assisted method according to claim 24, wherein: the information is extracted by steps of assigning at least one score to each sentence in the document according to the relevance of the sentence to the at least one dummy query, and wherein at least the highest scoring sentence is extracted.

29. The computer-assisted method according to claim 24, wherein: the information cached includes one document summary generated from the information extracted for each of the at least two dummy queries, and a label consisting of each term of the corresponding dummy query, and the summary generated consists of the document summary associated with the dummy query in which the terms of the label substantially match the terms of the user query.

Previous Doc Next Doc Go to Doc#

L4: Entry 6 of 8 File: USPT Mar 8, 2005

DOCUMENT-IDENTIFIER: US 6865572 B2
TITLE: Dynamically delivering, displaying document content as encapsulated within plurality of capsule overviews with topic stamp

Brief Summary Text (4):
Documents obtained via an electronic medium (i.e., the Internet or on-line services, such as AOL, Compuserve or other services) are often provided in such volume that it is important to be able to summarize them. Oftentimes, it is desired to be able to quickly obtain a brief (i.e., a few sentences or a paragraph length) summary of the document rather than reading it in its completeness. Most typically, such documents span several paragraphs to several pages in length. The present invention concerns itself with this kind of document, hereinafter referred to as average length document.

Brief Summary Text (5):
Present day summarization technologies fall short of delivering fully informative summaries of documents. To some extent, this is so because of shortcomings of the state-of-the-art in natural language processing; in general, the issue of how to customize a summarization procedure for a specific information seeking task is still an open one. However, given the rapidly growing volume of document-based information on-line, the need for any kind of document abstraction mechanism is so great that summarization technologies are beginning to get deployed in real world situations.

Brief Summary Text (8):
Given the extracted fragments which any particular method has identified as worth preserving, what is an optimal way of encapsulating these into a coherent whole, for presenting to the user? Acknowledging that different information management tasks may require different kinds of summary, even from the same document, how should the data discarded by the reduction process be retained, in case a reference is necessary to a part of the document not originally included in the summary? What are the trade-offs in fixing the granularity of analysis: for instance, are sentences better than paragraphs as information-bearing passages, or are phrases even better? Of particular importance to this invention is the question of "user involvement." From the end-user's point of view, making judgements, on the basis of a summary, concerning what a document is about and whether to pay it closer attention would engage the user in a sequence of actions: look at the summary, absorb its semantic impact, infer what the document might be about, decide whether to consult the source, somehow call up the full document, and navigate to the point(s) of interest. Given that this introduces a serious amount of cognitive and operational overhead, what are the implications for the user when they are faced with a large, and growing, number of documents to deal with on a daily basis?

Brief Summary Text (9):
These are only some of the questions concerning the acceptability of summarization technology by end users. There is particular urgency, given the currently evolving notion of "information push", where content arriving unsolicited, and in large quantities, at individual workstations threatens users with real and immediate information overload. To the extent that broad coverage summarization techniques are beginning to get deployed in real world situations, it is still the case that these techniques are based primarily on sentence extraction methods. In such a context, the above questions take on more specific interpretations. Thus, is it appropriate to concatenate together the sentences extracted as representative--especially when they come from disjoint parts of the source document? What could be done, within a sentence extraction framework, to ensure that all `themes` in a document get represented by the set of sentences identified by the technology? How can the jarring effect of `dangling` (and unresolved) references in the selection--without any obvious means of identifying the referents in the original text--be overcome? What mechanisms could be developed for offering the user additional

information from the document, for more focused attention to detail? What is the value of the sentence, as a basic information-bearing unit, as a window into a multi-document space?

Brief Summary Text (17):
Both examples present summaries as sentences which almost seamlessly follow one another. While this may account for acceptable readability, it is at best misleading, as in the original documents these sentences are several paragraphs apart. This makes it hard to know that the references to "How true, how true", in the first example, and The way this has turned out, in the second, are not whatever might be mentioned in the preceding summary sentences, but are, in fact, hidden somewhere in the original text of the documents. Opening references to "The problem", and "the agency", are hard to resolve. The thrust of the second article--namely that there is a reversal of an anticipated situation--is not at all captured: it turns out that the missing paragraphs between the summary sentences discuss how the planned 5,000 layoffs have been reduced to "4,000, then 1,400 and finally settled at about 500", and that "now, even those 500 workers will not be cut". As it turns out, some indication to this effect might have been surmised from the full title of the article, Scheduled IRS Layoffs For 500 Are Canceled, unfortunately, this has been truncated by a data reduction strategy which is insensitive to notions of linguistic phrases, auxiliary verb constructions, mood, and so forth.

Detailed Description Text (10):
The solution in accordance with the present invention to the problem of effectively communicating to the end user the `gist` of an on-line document, or of a collection of on-line documents, is based on the idea of relating form and content, by means of dynamic visual treatment of written language, or temporal typography. Only recently has the possibility of escaping the static and rigid constraints of writing on paper been fully appreciated. Wong, in Temporal Typography, Characterization of Time-Varying Typographic Forms (Master's thesis, MIT Media Lab, 1995), has stated: "Imagine looking at a small area on a computer screen. Words appear and disappear on the screen one by one. As they appear, meaning is expressed as forms change dynamically over time. The combined effect of the message, form and rhythm express a tone of voice, emotion or personality as if you hear a person speak. Although the two mediums, spoken and written words, are vastly different, the analogy may give you a sense of the expressive potential of temporal typography."

Detailed Description Text (11):
The notion, essentially, is to relate highlights of the core meaning of a message to ways of visually enhancing their impact, or at least mimicking (some of) their semantic load. In the immediate context of this disclosure, this translates to questions of what might be appropriate visual metaphors for representing semantic objects like topical phrases, shifts in discourse structure, or contextualization of information-bearing phrasal units.

Detailed Description Text (12):
There are several appealing aspects to dynamically presenting abstractions of document content. The user need not be actively involved: as documents arrive at the desktop, they can be analyzed and the resulting content abstractions can be displayed autonomously. Should the user have the time or inclination to focus on a particular document, interactive controls will be at their disposal; alternatively, each new arrival can be presented under its own schedule, followed by another, and so on. The presentation cycle can be customized to make use of arbitrary combinations of granularity of expressiveness. Notions like semantic highlights and demarcation of context are easily mapped onto visual metaphors, and thus naturally support the expression of content by means of variations of form. Cognitively, short phrases with high semantic load are amenable to punctuated display following a natural rhythm of visual perception.

Detailed Description Text (15):
collections of highly salient topical phrases,

Detailed Description Text (18):
organized as capsule overviews which track the occurrence of topical phrases and other discourse referents across the document discourse.

Detailed Description Text (22):
Through capsule overviews, a document's content is characterized in a way that is representative of the full flow of the document. This is in contrast to passage extraction

techniques, which typically highlight only certain fragments. Also, capsule overviews are derived by carrying out linguistically intensive analysis of the text in a document, which seeks semantic prominence of linguistics expressions, rather than just occurrence of certain pre-specified, or highly frequent, words and phrases--thus the system and method described here can be applied to any document, independent of domain, style or genre.

Detailed Description Text (23):
A capsule overview is not an instantiated template. A primary consideration of the content characterization system and method described here is that they should not be specific to any document source or type. A capsule overview is a coherently presented list of linguistic expressions which refer to the most prominent objects mentioned in the document, i.e., its topic stamps, and furthermore provide richer specification of the relational contexts (e.g., verb phrases, minimal clauses) in which these expressions appear.

Detailed Description Text (24):
To further illustrate the concepts associated with a capsule overview, refer now to the following news article shown in Table 1. (Marking certain phrase units within single quotes is an annotation device, for subsequent references to the text from within this disclosure document; these annotations were not part of the original article.)

Detailed Description Text (25):
There are a number of reasons why the title, "Priest Is Charged with Pope Attack", is a highly representative abstraction of the content of the passage. It encapsulates the essence of what the story is about: there are two actors, identified by their most prominent characteristics; one of them has been attacked by the other; the perpetrator has been charged; there is an implication of malice to the act. The title brings the complete set of salient facts together, in a thoughtfully composed statement, designed to be brief yet informative. Whether a present day natural language analysis program can derive--without being primed of a domain and genre--the information required to generate such a summary is arguable. (This is assuming, of course, that natural language generation techniques could, in their own right, do the planning and delivery of such a concise and information-packed message.) However, part of the task of delivering accurate content characterization is being able to identify the components of this abstraction (e.g., `priest`, `pope attack`, `charged with`). It is from these components that, eventually, a true summary of this document would begin to be constructed.

Detailed Description Text (28):
The core information unit that the invention concerns itself with is the set of discourse referents in a document. Discourse referents are typically realized as noun phrases. In essence, these are the entities--actors and objects--around which a story unfolds. In order to determine, and maintain, an accurate model of what a document is about, it is necessary to be able to identify the ways in which the same entity is referred to in the text, as well as to establish co-referentiality among different `mentions` in the text of the same entity. The sample document in Table 1 provides examples of the same entity being referred to in different ways in the text ("priest", "a Spanish Priest", "Fernandez", and "he", in the second paragraph, all refer to the same person), as well as of different entities being referred to by the same text string ("he" in the first paragraph refers to the Pope, while "he" in the second paragraph refers to the priest).

Detailed Description Text (31):
Salience is a measure of the relative prominence of objects in discourse: objects with high salience are the focus of attention; those with low salience are at the periphery. In an effort to resolve the problems facing a term-based approach to content characterization, as discussed in the background of the application, a procedure in accordance with the present invention has been developed which uses a salience feature as the basis for a "ranking by importance" of an unstructured referent set; ultimately, this facilitates topic stamp identification. By determining the salience of the members of a referent set, an ordering can be imposed which, in connection with an appropriate choice of threshold value, permits the reduction of the entire referent set to only those expressions that identify the most prominent participants in the discourse. This reduced set of terms, in combination with information about local context at various levels of granularity (verb phrase, minimal clause, sentence, etc.) offers an accurate and detailed characterization of a document's content. This may then be folded into an appropriate presentation metaphor such as that will be described hereinafter. Crucially, such an analysis satisfies some important requirements of usability of document content

abstractions: it is concise, it is coherent, and it does not introduce cognitive overload. In a more general sense, this method utilizes a strategy for scaling up the phrasal analysis techniques utilized by standard term identification and template instantiation technologies, which has at its core the utilization of a crucial feature of discourse structure: the prominence, over some segment of text, of particular referents--something that is missing from the traditional technology for `bare` terminology identification.

Detailed Description Text (33):
For the purposes of determining how discourse referents relate to objects in the world of the document, a simplifying assumption is made that every noun phrase identified by extended phrasal analysis constitutes a "mention" of a participant in the discourse. In order to determine which expressions constitute mentions of the same referent, the method described here crucially relies upon being able to carry out anaphora resolution and co-referent identification. Linguistic expressions that are identified as coreferential are grouped into equivalence classes, and each equivalence class is taken to represent a unique referent in the discourse. The set of such equivalence classes constitutes the fill referent set from which, ultimately, topic stamps will be derived.

Detailed Description Text (35):
The immediate result of anaphora resolution is to reduce the extended phrase set of all mentions of objects in the discourse; the larger consequence is that it provides the basis for the identification of topic stamps, as it introduces both a working definition of salience and a formal mechanism for determining the salience of particular linguistic expressions. This connection between anaphora resolution, co-reference identification, discourse salience, and semantic prominence is described in fuller detail in "Anaphora for Everyone: Pronominal Anaphora Resolution Without a Parser," (C. Kennedy and B. Boguraev, in Proceedings of COLING-96 (16th International Conference on Computational Linguistics), Copenhagen, D K, Aug. 5-9, 1996) and "Anaphora in a Wider Context: Tracking Discourse Referents" (C. Kennedy and B. Boguraev, in W. Wahlster, Editor, Proceedings of ECAI-96 (12th European Conference on Artificial Intelligence), Budapest, Hungary, Aug. 11-16, 1996. John Wiley and Sons, Ltd., London/New York).

Detailed Description Text (43):
The notion "segment of text" plays an extremely important role in the content characterization task, as it provides the basic information-structuring units around which a capsule overview for a document is constructed. Again, the example from Table 1 gives a useful illustration of the important issues. The reason that the title of this passage works as an overview of its content is because the text itself is fairly short. As a text increases in length, the "completeness" of a short description as a characterization of content deteriorates. If the intention is to use concise descriptions consisting of one or two salient phrases--i.e., topic stamps--along with information about the local context in which they appear as the primary information-bearing units for a capsule overview, then it follows that texts longer than a few paragraphs must be broken down into smaller units or "segments".

Detailed Description Text (46):
Striving to balance the conflicting requirements of depth and accuracy of a summary with those of domain- and genre-independence, the notion of a capsule overviews has been developed as content abstraction for text documents, explicitly designed to capture "aboutness". One of the problems of information management, when presented with a growing surplus of text documents, is getting some appreciation--rapidly, compactly, and yet with a usable degree of depth and representativeness--of the information contained in a document. Informally, this is usually referred to as the "aboutness" of a document, and is represented as a set of highly salient, and by that token most representative, phrases in the document. By viewing topicality in its stricter, linguistic, sense, the previous section defined topic stamps to be the most prominent of these phrases, introduced into, and then elaborated upon, the document body. On the basis of this definition, the above-identified computational, algorithmic, procedure has been developed for generating a set of abstractions for the core meaning in the document, ultimately resulting in a capsule overview of the document based upon suitable presentation of the most representative, and most contentful, expressions in the text. These abstractions comprise layered and inter-related phrasal units at different levels of granularity and depth of document analysis. To further describe this concept of granularity refer now to the following discussion.

Detailed Description Text (47):
Granularity is closely tied to context. In general, the information in a given sentence is best expanded by being able to position this sentence in its paragraph context; likewise, the theme and topic(s) in a paragraph can be further elaborated by relating the paragraph to the segment of discourse which encompasses the theme in its entirety. This is a natural containment hierarchy, relating the different information levels in a document together. Such a hierarchy can also be extended in sub-sentential direction: phrasal units indicative of topicality are clearly wholly contained in sentences; furthermore, a phrasal containment hierarchy could also be utilized to provide contextualized information concerning the topical phrases themselves.

Detailed Description Text (48):
Imagine that in the second example above (Example 2, page 5) some mechanism has determined that the phrase "Scheduled IRS Layoffs" is topically indicative. Assuming some focused mining in the vicinity of such an `anchor` by a phrasal grammar of a certain type, this topic phrase could be further contextualized to "Scheduled IRS Layoffs For 500 Are Canceled". This is an example of phrasal containment of information-bearing phrasal units. Similar expansion of topic in context might yield, for the initial discourse segment of the document, progressively larger and more informative fragments from it:

Detailed Description Text (50):
This example illustrates the notion of granularity of document analysis, and is especially indicative of how a containment hierarchy of layered information--from very compact and representative topical phrases all the way to full and rich discourse segments--can be utilized to represent and maintain strong notion of contextualization in a document abstraction.

Detailed Description Text (53):
It is clear that granularity of analysis and containment hierarchy of information-bearing phrasal units with different (yet complementary) discourse properties and function could be utilized very effectively to implement a "zooming" function into and/or out of a given document. In this way finding out more of what is behind a document "summary" is, in effect, filling in the gaps in such a summary in a controlled fashion, guided by incrementally revealing progressively larger and more informative contexts.

Detailed Description Text (58):
The following discussion describes an example of an article the analysis of which utilizes the present invention. As described in sections 2 and 3 above, the operational components of salience-based content characterization fall in the following categories: discourse segmentation; phrasal analysis (of nominal expressions and their relational contexts), anaphora resolution and generation of a referent set; calculation of discourse salience and identification of topic stamps; and enriching topic stamps with information about relational context(s). Some of the functionality follows directly from technology developed for the purposes of phrasal identification, suitably augmented with mechanisms for maintaining phrase containment; in particular, both relation identification and extended phrasal analysis are carried out by running a phrasal grammar over a stream of text tokens tagged for lexical, morphological, and syntactic information, and for grammatical function; this is in addition to a grammar mining for terms and, generally, referents.

Detailed Description Text (65):
The third segment (Table 5) of the passage exemplified above is associated with the stamps "Gilbert Amelio" and "new operating system". The reasons, and linguistic rationale, for the selection of these particular noun phrases as topical are essentially identical to the intuition behind "priest" and "Pope attack" being the central topics of the example in Table 1. The computational justification for the choices lies in the extremely high values of salience, resulting from taking into account a number of factors: co-referentiality between "amelio" and "Gilbert Amelio", co-referentiality between "amelio" and "his", syntactic prominence of "amelio" (as a subject) promoting topical status higher than for instance "Apple" (which appears in adjunct positions), high overall frequency (four, counting the anaphor, as opposed to three for "Apple"--even if the two get the same number of text occurrences in the segment), and boost in global salience measures, due to "priming" effects of both referents for "Gilbert Amelio" and "operating system" in the prior discourse of the two preceding segments. Compared to a single phrase summary in the form of, say, "Amelio seeks a new operating system", the overview for the closing segment comes close; arguably, it is even better than any single phrase summary.

Detailed Description Text (70):
Previously, the predominant current mechanism for mediating the spectrum between a summary of a document and a complete version of the same document was briefly discussed. In addition to a direct hypertext rendering of extracted sentences, in their full document contexts, two variations on this approach are the VESPA slider and HYPERGEN. VESPA is an experimental interface to Apple's sentence-based summarizer (Advanced Technologies Group, Apple Computer, Cupertino, Calif., Apple Information Access Toolkit: Developer Notes and APIs, 1997), whose main feature is a slider which dynamically readjusts the shrink factor of a document summary. HYPERGEN exploits notions of phrasal containment within sentence units, in an attempt to elaborate a notion similar to that of granularity of analysis and context introduced earlier in this document: in a process called sentence simplification, Mahesh (K. Mahesh, Hypertext summary extraction for fast document browsing, in Proceedings of AAAI Spring Symposium on Natural Language Processing for the World Wide Web, pages 95-104, Stanford, Calif., 19975) uses phrases as "sentence surrogates", which are then straightforwardly rendered as hypertext links to the sentences themselves.

Detailed Description Text (71):
As part of an ongoing investigation of visualizing large information spaces, researchers at Xerox PARC have looked at a variety of structured data types (such as hierarchically structured data, calendars, and bibliographic databases). Some general principles derived from that work have been applied to unstructured documents: the DOCUMENT LENS is a technique for viewing 2-D information, designed for component presentations of multi-page documents. Without going into detail, what is of particular relevance here is the strong notion of focus plus context which drives the design. The visualization, however, does little in terms of using any kind of document summary or other abstraction, and is of a predominantly static nature (even though it is extremely responsive to user interaction, as it attempts to combine a `bird's eye view` of the entire document with a page browsing metaphor). More recently experimental prototypes have been developed for interfaces which treat term sets (in the information retrieval sense, i.e. flat lists of index terms) as document surrogates: the focus of such designs is visually on presenting notions like distribution of terms across the document, and on mediating access to local context for a given term (R. Rao, J. O. Pedersen, M. A. Hearst, J. D. Macinlay, S. K. Card, L. Masinter, P. -K. Halvorsen, and G. G. Robertson, "Rich interaction in the digital library", Communication of the ACM, 38(4):29-39, 1995; M. A. Hearst, "Tilebars: Visualization of term distribution information in full text information access," in ACM SIGCHI Conference on Human Factors in Computing Systems, Denver, Colo., 1995). Ultimately, however, these interfaces still offer only a direct link between two states, the document surrogate and its fill form.

Detailed Description Text (74):
None of the examples above, however, combines a ticker with an automatic summarization engine. To a large extent this is because sentences--especially inconsecutive ones, in the absence of visual markers for discontinuity--do not lend themselves easily into the word by word, left to right, presentation mode. This is clearly a situation where phrasal units of a sub-sentence granularity can be utilized much more effectively. In addition, psychological experiments on active reading (Y. Y. Wong, Temporal typography, characterization of time-varying typographic forms, Master's thesis, MIT Media Lab, 1995) show that when text is presented dynamically in the manner of a ticker, subjects' reading speeds are significantly slower than for text presented statistically. On the other hand, dynamic presentations of text which show words or short phrases in the same location, but serially, one after the other, have reading speeds comparable to those for normal static texts.

Detailed Description Text (75):
To date, no applications have been developed utilizing temporal typography for dynamic delivery of content abstractions. Wong has looked at how dynamic type in general can be used for four different communicative goals: expressive messages, dialogue, active reading and real time conversation. Most relevant to this discussion are her experiments on active reading. In one of these she used a basic RSVP (Rapid Serial Visual Presentation) method (words or phrases presented sequentially one after another, on the same line and at the same position) to deliver a sequence of news headlines. In a second set of experiments called HIGHWAY NEWS, three dimensions are utilized, combined with a zooming motion, to present a sequence of text highlights. "News headlines are placed one after another in the z-dimension. Headlines are presented serially according to active input from the reader presses a mouse button to fly through the rows of headlines--as if flying over highway of text." These experiments show the

strong feasibility of high impact, low engagement, delivery of semantically prominent text fragments being utilized as a powerful technique for visualizing certain types of inherently linear information.

Detailed Description Text (86):
In its basic mode with no user interaction, the RSVP Viewer cycles through all salient relational contexts in a document, maintaining the order in which they appear in the text. As shown in FIGS. 7A-7B, each context phrase is displayed as the prominent object on the screen 702; at the same time the context is overlaid onto topic expansions (displayed as translucent text) 704. This facilitates further interpretations of the context strings by the user: expansions relate phrasal contractions in the string displayed to their full canonical forms in the text, make clear antecedents of dangling anaphors, and so forth. Note, for instance, the background display (FIG. 7C) 704 of the full form of the antecedent for the anaphoric "he" in the foreground 702: in this particular context, "he" has been resolved to "Gilbert Armelio".

Detailed Description Text (87):
Cycling though the complete set of salient contexts, in their original sequence, offers a good indication of aboutness at a given level of depth and detail. Granularity of display is adjustable via a parameter: thus RSVP Viewer could be reduced to a TopicsTicker Viewer by only cycling through the document's topic stamps, or it could be used to display sequences of sentences. Relational contexts offer just the right balance between terseness (phrases are more easily perceived and assimilated than sentences) and informativeness (phrases larger than `bare` topic stamps convey richer data). The amount of time a phrase is displayed is dynamically calculated, based on studies of active reading and perception; the intent is to optimize the full document display regime so that individually, each phrase can be processed by the user, while globally, the entire set of salient contexts can be cycled through rapidly.

Detailed Description Text (95):
The ViewTool viewer places the capsule overview of a document within the context of the document itself. This is maintained by synchronized display of discourse segments, topic stamps, and relational contexts in three panels. The whole document 816 is displayed in the left panel; this is deliberately unreadable, and is intended to function as a document thumbnail serving as a contextual referent for the topics presented in the central panel. With the use of an appropriate color coding scheme, it also serves as an indicator of the distribution of topically prominent phrases in the document. The central panel 804 lists the highly salient topic stamps. Contextualization for these is achieved by aligning the topic stamps for a given discourse segment with the textual span of that segment in the thumbnail as indicated by 808 and 810 in FIG. 8A and discussed below. This offers an immediate overview of, for instance, what is being discussed in the beginning of the document, or in the end, or which topics keep recurring throughout, and so forth.

Detailed Description Text (96):
The central panel is sensitive to the user's focus of attention: as the mouse rolls over a topic stamp 804, the discourse segment from which this topic has been extracted is highlighted in the left panel shown at 810. The highlighting also indicates the segmentation of the source documents into topically different, and distinct, text sections. This design makes it easy to do rapid selection of areas of interest in the document, as it is mediated by the topic stamps per segment display. Again, the granularity of analysis and the layered contextual information in the capsule overview make it easy to offer immediate and more detailed information about any given set of topic stamps: simultaneously with highlighting the appropriate discourse segment 810 in the left panel 802, relational contexts for the same set of topic stamps 808 and 812 are displayed cyclically, in RSVP-like fashion 814, in the right panel 806. This ensures that topic stamps are always related with contextual cue phrases. Thus an additional level of detail is made available to the user, with very little `prompting` on their part. On the other hand, as FIG. 8B illustrates, if it is still the case that the full text of the segment would be required, clicking on its `proxy` topic stamps 808 (in the middle panel) would display this in the right panel 818. The larger area available there, as well as an automatic readjustment of the size of type, ensures that the text is readable.

Detailed Description Text (97):
Referring now to FIG. 8C, as a natural extension of the same metaphor, clicking on the document proxy 816 in the left panel brings up the full document text in the right panel 820. The full text always uses color markup to indicate, in yet another way, topically salient phrases and

their relational contexts.

Detailed Description Paragraph Table (3):
TABLE 2 "sent": 100 iff the expression is in the current sentence. "cntx": 50 iff the expression is in the current discourse segment. "subj": 80 iff the expression is a subject. "exst": 70 iff the expression is in an existential construction. "poss": 65 iff the expression is a possessive. "acc": 50 iff the expression is a direct object. "dat": 40 iff the expression is an indirect object. "oblq": 30 iff the expression is the complement of a preposition. "head": 80 iff the expression is not contained in another phrase. "arg": 50 iff the expression is not contained in an adjunct.

Detailed Description Paragraph Table (4):
TOPICAL PHRASE: "Scheduled IRS Layoffs" TOPIC IN RELATIONAL "there will be no layoffs" CONTEXT: TOPICAL SENTENCE: "Yesterday, the IRS said there will be no layoffs" SENTENCE IN PARAGRAPH "More than a year ago, CONTEXT: The Internal Revenue Service planned widespread job cuts. Yesterday, the IRS said there will be no layoffs." PARAGRAPH WITHIN TOPICALLY "More than a year ago, COHERENT DISCOURSE THEME: the Internal Revenue Service planned widespread job cuts. Yesterday, the IRS said there will be no layoffs." Confronted with congressional criticism and calls for reform in light of some highly publi- cized reports of abusive actions toward taxpayers, as well as staunch union opposition to the cuts, the IRS said employees at risk of losing their jobs would be reassigned to improve `customer service,` help taxpayers resolve problems and increase compliance with tax laws."

Previous Doc Next Doc Go to Doc#

L8: Entry 5 of 5 File: USPT Aug 1, 2000

DOCUMENT-IDENTIFIER: US 6098034 A
TITLE: Method for standardizing phrasing in a document

Abstract Text (1):
A method for standardizing phrases in a document includes the steps of identifying phrases of a document to create a preliminary list of standard phrases; filtering the preliminary list of standard phrases to create a final list of standard phrases; identifying candidate phrases of the document which are similar to the standard phrases; confirming whether a candidate phrase of the document is sufficiently proximate to the standard phrase to constitute an approximate phrase; and computing a phrase substitution to determine the appropriate conformation of standard phrase to the approximate phrase or the approximate phrase to the standard. Further this invention relates to a computer system for standardizing a document.

Brief Summary Text (2):
The problem addressed by this application is the identification in a document of user significant phrases, as indicated by repetition of particular sequences of words. Typically, in documents created by precision-oriented users of various sorts, certain groups of words are used to convey a particular idea. Each time the user desires to express that idea, the user prefers to utilize the same phrasing over others in order to avoid confusion of meaning. In order to determine whether a significant phrase has already been used in the documents, the user must review the document to extract the relevant sequence of words. Once a phrase has been extracted, the user may refer to it on a continual basis to ensure that, whenever that user desires to express a similar idea, a form of the extracted sequence of words is used.

Brief Summary Text (4):
The problem of standardizing phrasing, as described above, is one currently performed only manually. The human user conducts a time-consuming review of a document for significant phrases. This review is made in an attempt to detect the standard way of phrasing an idea in order to ensure continued phrasing of that idea in a manner that conforms to earlier phrasing.

Brief Summary Text (5):
Further, the human reviewer seeks to identify similar yet non-identical phrases in order to conform them. There is generally no explicit extraction and designation of standard phrases; these phrases are left within their contexts and simply used as the standards to which similar

Brief Summary Text (8):
In addition, the suffix tree is usually used for the applications of storage, compression, and searching. In the subject application, the tree is used not for document or phrase storage, but rather for phrase identification by establishing word sequences that satisfy the criteria for length and recurrence in the document. In more detail, each node of the tree is associated with a record of the number of occurrences of the word sequence at that node; any such word sequence of sufficient length, where the number of occurrences exceeds the required threshold, is preliminarily designated a phrase. Inclusion on the final phrase list follows the post-processing steps outlined below.

Brief Summary Text (10):
An algorithm for the construction of a word-based suffix tree has been published by Andersson, et al. (Andersson, A. Larsson Jesper N. Swanson, K. "Suffix Tree on Words," Department of Computer Science, Lund University, Jul. 12, 1995.) Andersson, et al. is neither related to nor contains aspects related to the subject invention because Andersson does not relate at all to the overall process that is the subject of this application, standardizing document phrasing. Further, Andersson deals only with the construction of a word-level suffix tree; it does not

relate at all to the process of standard phrase extraction. Further, Andersson constructs its word-based suffix tree on the level of the entire document and does not innovate the sentence suffix tree structure that enables the subject method its unique combination of efficiency and non-complexity. Further, Andersson does not attempt to pre-process the text at all through stemming and abstraction of known characters. Lastly, Andersson, does not address the related problem of nested phrases or any resolution thereof.

Brief Summary Text (12):
The subject invention is a method for standardizing user phrasing in a user-created document. This method involves two separate components, each of which is further divided into separate steps. The first component involves the automatic extraction from the document of sequences of words constituting significant user phrases. This is accomplished through the combination of document pre-processing, the representation and analysis of the document text on a modified suffix tree, and heuristic post-processing. The second component involves the automatic extraction from the document of sequences of words that are significantly similar but not identical to significant user phrases, and the automatic generation of suggested phrasing for this approximately matched phrasing that conforms its phrasing to the standard. This is accomplished through a combination of the location of candidate word sequences, a computation of the weighted "edit distance" between the significant phrase and the phrase approximately similar to it, and certain natural language processing techniques. The overall result of this method is a list of significant user-created standard phrases and the standardization of approximately matched phrasing throughout the document.

Drawing Description Text (2):
FIG. 1: Process data flow diagram of process of standardizing a document.

Drawing Description Text (8):
FIG. 7: Block diagram of computer system for standardizing the phrasing of a document.

Detailed Description Text (2):
The present invention relates to a method for standardizing phrases in a document, which comprises the steps of: identifying phrases of a document to create a preliminary list of standard phrases; filtering the preliminary list of standard phrases to create a final list of standard phrases; identifying candidate phrases of the document which are similar to the standard phrases; confirming whether a candidate phrase of the document is sufficiently proximate to the standard phrase to constitute an approximate phrase; and computing a phrase substitution to determine the appropriate conformation of standard phrase to the approximate phrase or the approximate phrase to the standard.

Detailed Description Text (3):
In one embodiment the step of identifying phrases of a document to create a preliminary list of standard phrases further comprises tokenizing the document. In another embodiment of the invention the step of identifying phrases of a document to create a preliminary list of standard phrases further comprises constructing sentence suffix trees for determination of standard phrases. In another embodiment of the invention the step of identifying phrases of a document to create a preliminary list of standard phrases further comprises traversing and analyzing the suffix tree. In another embodiment the step of identifying phrases of a document to create a preliminary list of standard phrases further comprises the application of a stop list.

Detailed Description Text (4):
The high level design of the first component of the process, the automatic extraction of standard, user phrases, is centered around the construction of a modified suffix tree in which stemmed words are the atomic units; this tree is traversed and analyzed to identify sequences of words meeting certain pre-set criteria for significant phrases. The use of the suffix tree is supplemented by several additional steps that, together, enhance the accuracy and efficiency of the process, and hence the number of standard phrases found. First, the processes of stemming, application of a stop list, and abstraction for known structural elements integrate information retrieval and other techniques to ignore elements of the document that do not affect the overall meaning of the sequence of words. These processes thus broaden the candidate group of phrases. Second, effectiveness is further enhanced by the post-construction processes of eliminating nested sub-phrases that do not occur sufficiently frequently independent of their nesting phrases. Finally, validity of the extracted phrase is supplemented by eliminating

dangling minimal content words.

Detailed Description Text (6):
Certain processing is conducted in connection with the suffix tree in order to heighten the accuracy and efficiency of the phrase finding. First, there are certain structural elements in a document, detailed herein, that are known to be structural in nature and do not affect the meaning of the surrounding sequences of words. As these words do not serve to supplement or modify the meaning of the sequences of words preceding or following them, they are abstracted in the construction of the suffix tree; at any point where a given type of structural element appears, a marker is inserted to indicate the type of element and to hold the element's place. This enables the process to treat sequences of words, similar in all respects other than the presence of these known structural elements, as standard phrases. This improves the inclusiveness and accuracy of the results.

Detailed Description Text (13):
In one embodiment the step of identifying candidate phrases of the document which are similar to the standard phrases of the above invention the standardizing method further comprises constructing of a dictionary and phrase index. In another embodiment of the subjetc invention the step of identifying candidate phrases of the document which are similar to the standard phrases further comprises identifying candidate phrases by searching for approximate matches. In another embodiment of the subject invention the step of identifying candidate phrases of the document which are similar to the standard phrases further comprises application of a shifting window. In another embodiment of the subject invention the step of identifying candidate phrases of the document which are similar to the standard phrases further comprises accumulation of counts. In another embodiment of the subject invention the step of identifying candidate phrases of the document which are similar to the standard phrases further comprises generating candidates.

Detailed Description Text (14):
The dictionary of standard phrases that forms the basis of the determination of substantial similarity, termed the "Dictionary," is broken down into two coordinating data structures. These structures include both a list of phrases and a character-based "Phrase Index" containing all the characters, stemmed words and locations of these stemmed words contained in the list of phrases. This dual-structured dictionary enables efficiency in comparing the text under analysis to the standard phrases and, in particular, filtration of the document in order to locate candidate word sequences that may be substantially similar to a user-created standard phrase. These candidate phrases are not necessarily substantially similar to the standard; they simply meet certain word usage criteria that make them worthy of further analysis via the edit distance calculation. As the edit distance computation is time-consuming, this filtration phase is necessary in order to markedly reduce the number of phrases for which edit distance is to be calculated. This finding of candidate phrases is accomplished through the traversal of the text by a "sliding window," seeking units of text containing certain words in common with the words contained in the dictionary of standard phrases. Where a sufficiently large number of words is found in common, irrespective of the order of their appearance, the sequence of words is designated a candidate phrase, deserving of further analysis by the time-intensive edit distance calculation.

Detailed Description Text (18):
of either one of the two approximately similar phrases, one the user-created standard phrase, the other the phrase substantially similar to it, may be desired. Regardless of the option selected, only the discrepancies between the two phrases are transformed, while the remainder of the attributes and content of the conformed phrase are retained, in order to ensure the syntactic coherence of the document.

Detailed Description Text (20):
In one embodiment of the step of confirming whether a candidate phrase of the document is sufficiently proximate to the standard phrase to constitute an approximate phrase the standardizing method further comprises calculating the edit distance of candidate phrases. In another embodiment of the invention the standardizing method further comprises calculating the edit distance of candidate phrases. In another embodiment of the invention the standardizing method further comprises substituting conforming phrases. In another embodiment of the invention the standardizing method further comprises weighting the edit distance.

Detailed Description Text (21):
In one embodiment of the step of computing a phrase substitution to determine the appropriate conformation of standard phrase to the approximate phrase or the approximate phrase to the standard further comprises conforming the document for grammer. In another embodiment the standardizing method further comprises recommending a phrase that conforms the candidate phrase to the user standard phrase , while otherwise retaining the syntactic and grammatical coherence of the candidate phrase. In another embodiment the standardizing method further comprises computing a recommended phrase that conforms the user standard phrase to the candidate phrase, while otherwise retaining the syntactic and grammatical coherence of the candidate phrase.

Detailed Description Text (22):
In addition, in another embodiment of the above invention the standardizing method further comprises presenting a user with recommended phrases computed for editing and user approved insertion in the document.

Detailed Description Text (23):
"Abstraction" is defined as the process of replacing a sequence of words that constitute a known structural element with a marker representing the type of that element. "Approximate phrase" is defined as a sequence of words in a document whose edit distance from a known master phrase is below the given edit distance threshold. "Attribute" is defined as a category label associated with a word. "Canonical form" is defined as the standardized, "dictionary entry" form of a word, created by removing grammatical appendages such as conjugations or declensions. For example, the canonical form of "sold" is "sell" and that of "children" is "child".

Detailed Description Text (24):
Further, "conform" is defined as the process of editing an approximate phrase to correspond to its matching master phrase, or vice versa, retaining the grammatical structure of the source phrase as much as possible. "Conformed phrase occurrence" is defined as the word sequence resulting from automatically conforming an approximate phrase or a user standard phrase. "Domain" is defined as a particular field of discourse having its own specialized terminology and types of documents, e.g., law, software engineering, etc. "Edit distance" is defined as a measure of the distance between two sequences of words, consisting of a count of the number of word insertions, deletions, and replacements needed to transform one sequence to the other. "Edit distance threshold" is defined as a number giving the largest edit distance between a master phrase and another phrase, such that the second phrase will be considered an approximate phrase for the master phrase.

Detailed Description Text (25):
Further, "encoding" is defined as the process of transforming a sequence of words into a sequence of numbers, where each number represents a particular word (the word's code). "Lexical attribute" is defined as an attribute indicating if a particular word has independent meaning, or if its meaning is only in its grammatical function. "Master phrase" is defined as a word sequence constituting a standardized phrasing for a document. "Nested phrase" is defined as a pair of phrases P and P', such that the sequence of words in P' occurs as a subsequence of the sequence of words in P. P' is nested in P. Trie"is defined as a tree, each of whose nodes is labeled by a sequence of elements (e.g., characters or words), where each node represents the string formed by concatenating the labels of its ancestors starting with the root and ending with the label of the node itself". "Path compressed trie" is defined as a trie where no node has exactly one child. "Phrase" is defined as a sequence of words in canonical form constituting phrasing. "Prefix nested phrase" is defined as a nested phrase where the shorter phrase's word sequence is a prefix of the longer phrase's word sequence. "Segmentation" is defined as the process of converting a document represented as a sequence of characters into a sequence of sentences, each comprising a sequence of words.

Detailed Description Text (26):
Further, "semantic attribute" is defined as an attribute categorizing a sequence of words according to its meaning. For example, "Ab Def Inc." may be categorized as a "company name". Semantic categories may be organized in a hierarchy, with higher-level categories including many lower-level categories; a word is considered to have all semantic attributes whose categories it is included in. Note that in contradistinction to lexical attributes, semantic attributes may apply to sequences of several words. "Sentence suffix tree" is defined as a trie representing the set of sequences comprising all of the Suffixes of all of the sentences in a given document. "Structural element" is defined as a word, or sequence of words, in a document

that serve a unified functional purpose in the document, independent of the words' individual meanings. In a particular domain, different types of structural elements will exist. For example, in the domain of legal documents, one type of structural element is the "defined term", corresponding to usage of a term defined elsewhere in the document. Identifying structural elements and their types in a document may be done either manually or automatically, through the use of domain-dependent methods. "Stemming" is defined as the process of reducing a word to its canonical form. "Stop list" is defined as a list of frequently used words which bear minimal content on their own (such as conjunctions or prepositions). "Suffix tree" is defined as path compressed trie which represents exactly the set of sequences comprising the distinct suffixes of a given sequence of elements.

Detailed Description Text (27):
Further, "syntactic attribute" is defined as an attribute giving a word's syntactic function. For example, in English text, this may be as simple as a part-of-speech label ("noun", "verb", etc.) or a more complex descriptor based on the word's relationship to other surrounding words. "Template replacement" is defined as a method of the automatic conforming, where for a certain type of words, the approximate phrase's word is retained and not replaced by the corresponding word in a master phrase (for example conjunctions--"and" cannot be replaced by "or"). "Tokenization" is defined as the process of converting a document into a sequence of sentences, including the processes of segmentation, abstraction, stemming, and encoding. "Weighted edit distance" is defined as a variety of edit distance where the operations of insertion, deletion, and replacement are weighted differently, possibly according to the attributes of the words being compared.

Detailed Description Text (28):
In addition, the present invention relates to a computer system for standardizing a document, comprising a computer containing an executable standardizing document program in which such program imparts conforming functionality to the computer by changing the state of the computers logic unit upon execution.In another embodiment the computer program is stored on the computer hardware, software or in RAM. In the preferred embodiment the computer program is the process of standardizing a document as described herein.

Detailed Description Text (29):
In addition, the present invention relates to a computer system for standardizing a document, comprising: a. a display; b. a memory; c. a microprocessor; d. a storage; e. an input circuit; and f. an output circuit.

Detailed Description Text (35):
In more detail, the Master Phrase Finding Phase 2, detailed in FIG. 2 and below, itself consists of several steps. First, the pre-processing step involves the tokenization 12 of the document into sentence-based "word packs", the abstraction of known structural elements within the document, and the stemming of word packs. Second, the determination of phrases involves the construction of the suffix tree, the traversal and analysis of this tree, and the application of a "stop list" of words during this analysis. The output of this Phase is a preliminary list of user-specific phrases.

Detailed Description Text (39):
The Phrase Conformation Phase 10, detailed in FIG. 3 and below, then uses this sequence of editing operations to compute a recommended candidate phrase that conforms discrepancies between two phrases, while retaining the grammatical and semantic coherence of the target phrase. This is done by applying the editing operations in order, while applying syntactic and semantic substitution rules. The output is a Conformed Phrase Occurrence which approximates the correct substitution for the target phrase. The user is then given the option to edit the Conformed Phrase Occurrence before substituting it into the document text.

Detailed Description Text (40):
The result of the overall process is the standardization of the phrasing in a user-created document according to the user's own, phrasing choices.

Detailed Description Text (46):
source sequence is divided into a set of "sentences", a "sentence suffix tree" is a trie whose set of sequences represented are all of the suffixes of each sentence in the sequence.

Detailed Description Text (47):
In the subject application, the sequences represented are either sequences of words constituting documents or parts of a document, or sequences of characters, representing individual words.

Detailed Description Text (49):
As depicted in FIG. 2, the Master Phrase Finding Phase 2 consists of three major steps, each detailed in this section. These include (a) tokenization, detailed in FIG. 6, (comprising segmentation, abstraction, stemming, and encoding), (b) construction of the Sentence Suffix Tree 15 (the "SST"), and (c) traversal and analysis of the SST, including application of the stop list. These steps together result in the construction of the candidate library of user phrases.

Detailed Description Text (52):
Segmentation of the source text involves its deconstruction into units, termed "word packs", each one representing a word. Segmentation is performed on a sentence-by-sentence basis. Upon reaching the end of a sentence, the word packs from that sentence are ready for the next stage of tokenization. The process breaks up the source text into both words and sentences in parallel, extracting words one by one until a sentence boundary is reached. A standard finite-state machine technique is used to recognize patterns that indicate word and sentence boundaries. This method is efficient, extensible, and simple.

Detailed Description Text (53):
As an example of segmentation, consider the following (schematic) sample sentence, in which each of the letters below represents a word in that sentence. Further, assume that the word represented by the letter "a" below has different conjugated instances, as indicated below:

Detailed Description Text (55):
Segmentation divides the input characters in the source text into the sequence of words constituting this sentence, as a sequence of its component words. Once the complete sentence is produced, it is abstracted, as described below.

Detailed Description Text (57):
There are a variety of elements in a document that are known to be structural in nature and do not affect the meaning of the surrounding sequences of words. These elements include many different types of terms, including without limitation defined terms, references to certain sections, references to attachments and appendices, titles of related documents, names of persons party to or otherwise connected with the document and related documents, laws, rules, regulations, and many other proper nouns. None of these words serve to modify the meaning of the sequences of words preceding or following them. Thus, an analysis of a group of words containing any of these structural elements should proceed on the basis of, and is made more accurate by, normalizing these terms. Any term of a given type is functionally similar to any other term of the same type. Insertion of an abstraction marker for each type of structural element (indicating the position and type of the structural element) prior to the construction of the SST allows the analysis of the remainder of the words in the word group, whether these words are contiguous or are separated by the structural element. The abstraction process described here assumes that some other domain-dependent process has previously been applied to the source text, and that, for each type of structural element, a special "bracketing" character has been inserted at the start and at the end of each word sequence constituting an occurrence of that element type.

Detailed Description Text (58):
For example, assume the group of word packs tokenized from the above-provided sample sentence:

Detailed Description Text (70):
Once the sentence has been tokenized, abstracted, and stemmed, each word is encoded as a unique integer. The encoding is done in order to reduce the cost of comparing words to one another in the suffix tree construction and analysis phases, as well as during phrase filtering. Encoding is accomplished by representing each word in a trie, as described in Section IV(2), above. Each word is associated with a unique integer. When a new word is encountered, if it is found to appear already in the trie, the integer associated with that word is used as its code. Otherwise, the word is inserted into the trie and is assigned a new code. An ancillary array is also maintained, which indexes nodes in the trie corresponding to words according to their

assigned codes. The trie together with this "code index" is termed the "Word Index". Special markers, such as those associated with structural element abstractions, have predefined integer codes that are inserted directly. The sequence of words is thus converted to a sequence of integers, enabling words to be most efficiently compared to one another.

Detailed Description Text (71):
The set of word packs has now been completely pre-processed and is ready to be represented by, and analyzed in, the Sentence Suffix Tree.

Detailed Description Text (73):
A. Construction of the Sentence Suffix Tree

Detailed Description Text (74):
The sentence suffix tree utilized as a part of the overall process described herein is a specialized form of suffix tree, as noted in Section IV(1), above. In essence, a suffix tree for a given element is a method of representing every distinct sub-sequence of items constituting a suffix of that element. Were the general form of suffix tree utilized in the context of a document, the suffixes that the suffix tree would represent would be those of the entire document. In this context, the atomic units represented at each node on the tree would be stemmed word packs, not characters or character strings. Each leaf of the resultant suffix tree would represent a distinct suffix of the document as a whole. The use of the SST, as further detailed below, yields a different and more efficient result.

Detailed Description Text (80):
n=the number of elements in the document

Detailed Description Text (81):
This equation indicates that the time for construction using Q Method is a quadratic function of the number of words in the document. Because the suffix tree was created to handle problems that could be handled more efficiently, Q Method suffix trees were never known to be widely used. The use of suffix trees became practical and, in certain applications, widespread, as the methods for construction become more efficient. These methods include several methods, including that innovated by P. Weiner (Weiner, P. "Linear Pattern Matching Algorithms," Conference Record, IEEE 14th Annual Symposium on Switching and Automata Theory, pp. 1-11.) and improved by E. M. McCreight (McCreight, E. M., "A Space-Economical Suffix Tree Construction Algorithm," Journal of the Association for Computing Machinery, Vol. 23, No. 2, April 1976, pp. 262-272.), as well as that developed by Ukkonen (Ukkonen, E. "On-Line Construction of Suffix Trees," Algorithmica, Vol. 14, No. 3, September 1995, pp. 249-260) ,that for the purposes of this application, are termed collectively the L Methods. These methods are much more efficient than the Q Method, as they are able to construct the tree in linear time. That is, the amount of time necessary to construct the tree is proportional to the number of elements under analysis. In the case of a document, this implies that the time to construct a suffix tree for the document as a whole would be proportional to the number of words in the document. Using the L Method, the construction time is given by:

Detailed Description Text (82):
indicating the time for construction to be a linear function of the number of words in the document. However, the L method involves complex algorithms, difficult to implement and maintain. Therefore, the L Method was also deemed inappropriate for the purposes of the method that is the subject of this application.

Detailed Description Text (85):
This unique combination is enabled by the specific nature of the elements under analysis--documents--and the atomic units chosen--words. The nature of the document permits a significant and innovative modification in the method described herein, termed the EE Method. The significant phrases sought for identification in a document cannot by definition span a sentence boundary; if a whole sentence is a standard phrasing, then that sentence becomes designated a phrase. As a result, the Sentence Suffix Tree, constructed as part of this application, represents not the distinct sub-sequences of words in the document as a whole, but rather only the distinct sub-sequences of words in each individual sentence. This markedly reduces the time for construction of this tree to:

Detailed Description Text (87):

s=the number of words in a given sentence in the document

Detailed Description Text (88):
f=the number of sentences in the document

Detailed Description Text (89):
To simplify the equation, in the worst case scenario in a document with X sentences of length less than or equal to s.sub.max, the maximum time (T) that it takes to build a suffix tree is represented by the following formula:

Detailed Description Text (95):
X=1000 The number of sentences in a document;

Detailed Description Text (96):
s.sub.max =10 The number of words in each sentence in the document.

Detailed Description Text (101):
The EE Method is sufficiently efficient to allow construction of a Sentence Suffix Tree in near-linear time, yet accomplishes this with the far lesser complexity associated with the Q Method. This optimal combination of efficiency and complexity is highlighted as increasingly large documents are examined. It is important to recognize that, as documents grow, their number of sentences (A) grows, but the length of a sentence remains constant (s.sub.max). Therefore, as the only quadratic function in the EE Method is sentence length, which does not change with the length of the document, the time in the EE Method increases only linearly as the document grows in size.

Detailed Description Text (103):
As noted above, the SST constructed using the EE Method represents every distinct suffix subsequence of words. Each word sequence (suffix of a sentence) is inserted individually into the tree as follows (see the detailed flow-chart in FIG. 5). First, the tree is traversed from the root down in order to find the node with the longest corresponding word sequence which is a prefix of the current sequence (that to be inserted). If the node's sequence completes a prefix of the current sequence, a new child is added to that node, which child is labeled with the remaining suffix of the current sequence. Otherwise, the sequence corresponding to the node contains words beyond the corresponding prefix, and hence the node N must be split. This is done by creating a new node M, labeled with the first part of N's label (that contained in the prefix of the current sequence). N then becomes a child of M, and is labeled with the part of its old label remaining after M's label is removed. Then, the remaining suffix of the current sequence is inserted as a child of M, just as in the case where no splitting was required.

Detailed Description Text (107):
The tree is constructed iteratively, starting from the first term in the sentence and moving forward. Considering the suffix of the first term, the tree would appear as follows:

Detailed Description Text (113):
The tree at this point in the sentence would appear as follows:

Detailed Description Text (116):
Each node in this final sample tree represents distinct sub-sequences of words within the sentence, while each leaf represents a distinct suffix of a sentence. Beside each node is a number that represents the number of recurrences R of the particular word or set of words. The number of leaves is, at most, the number of words in the sentence. As a result, the maximal size of the tree is proportional to the size of the text as a whole.

Detailed Description Text (117):
As increasing numbers of sentences are examined by the EE Method, only incremental suffixes are added to the tree. Word sets and suffixes already contained on the tree are not added. Instead, a counter of the number of recurrences, R, is incremented.

Detailed Description Text (122):
L and R dictate the size of the phrase list found. In a typical document, as either parameter grows, the sets of words that qualify decrease. Where both these criteria are met, the set of words is determined, on a preliminary level, to be a significant phrase. The satisfaction of

both these criteria is apparent upon traversal of the EE tree constructed. Where the number recording recurrence at each node is below the threshold r.sub.min the word set ending at that node, and all word sets below it on that branch, do not have sufficient recurrence to satisfy the criteria for significance as a phrase. Where the number of words at a given node and above it on that branch are not sufficient to satisfy the length criteria l.sub.min, the word set ending at that node does not satisfy the criteria for significance as a phrase.

Detailed Description Text (204):
for two purposes. First, the phrase list itself may be output to the user interface for the user to refer to for the purpose of standardizing later usages of similar phrases to the standard phrasing identified by the steps outlined earlier. Second, the phrase list is further processed in the following phases in order to facilitate a further standardization of phrasing. In particular, the final phrase list is utilized in the construction of the phrase dictionary that forms the basis of the identification of sets of words that are approximately similar, but not identical, to the phrases. This process is detailed below.

Detailed Description Text (207):
The Final Phrase List generated above constitutes a list of phrases that are identical in all respects or near-identical, differing only in their usage of known structural elements. Each of the phrases contained on the Final Phrase List, as well as certain others that may be added manually, as described below, is termed a "Master Phrase." The second portion of the subject method seeks to identify phrases that differ to a more significant extent from a Master Phrase. These phrases, termed "Approximate Phrases," are substantially similar to one or more Master Phrases, yet may contain additional words, omit certain words, or replace certain words with other words. The process of identifying approximate matches involves two separate phases, the Candidate Phrase Finding Phase and the Candidate Phrase Confirmation Phase, each detailed in the sections below.

Detailed Description Text (215):
(III). Identification of Candidate Phrases via Traversal of the Document

Detailed Description Text (217):
Searching directly for approximate matches in the entire document would be extremely inefficient, as the time required would be approximately proportional to the product of the size of the document and the total size of the Dictionary. Instead, to avoid the extremely time-consuming process of seeking approximately matching word sequences throughout a document, the subject method utilizes two steps to identify Approximate Phrases. The first step, Candidate Phrase Finding, seeks to quickly eliminate most of the word sequences in the document that could not qualify as Approximate Phrases; what remains at this point are only those candidate phrases that could potentially be sufficiently similar to the Master Phrases to constitute Approximate Phrases. The objective at this point is to identify those phrases that may qualify as Approximate Phrases in order to further process each and so conclude whether they are Approximate Phrases. The second phase constitutes the further processing of these candidate phrases to determine whether they are sufficiently close in wording to the Master Phrases to qualify for designation as Approximate Phrases.

Detailed Description Text (231):
This process incrementally keeps track of the number of occurrences of words from the current window that are located in any relevant Master Phrase. This incremental approach enables great efficiency, since at any new location of the window, there need not be a comparison of the entire windowed text with the Phrase Index. As a result, the cost of the method increases only linearly with the size of the document and is not affected at all by the size of the window.

Detailed Description Text (233):
A further efficiency gain may be realized in this step by comparing only the first n letters in the word to the WPL's in the Phrase Index. If those n letters match, even if it is unknown whether the remainder of the word matches, the word is considered to match with a WPL. This is possible without introducing errors, since this step in the subject method is oriented not to finding definitive Approximate Phrases, but rather only candidate approximate matches deserving of further processing via edit distance calculation. Determining how many characters to match in order to optimize processing speed is done by empirical testing on typical candidate documents.

Detailed Description Text (247):
Second, edit distance is not used in the subject method to recommend replacement of the object under analysis with a particular objectively correct item, e.g., the correct spelling of a word. In the context of phrasing there is not necessarily an objectively correct phrase. In theory, any phrasing might be correct and it is the user's preference that dictates the style chosen. As a result, the process of extracting approximately-similar phrases is oriented toward presenting the user with user-defined phrasing options for similar phrases, in which the user may opt to modify either phrase to suit the other, whether the Approximate Phrase to suit the Master, or the Master Phrase to suit the Approximate. Regardless of the option chosen, a similar amount of the original Phrase is retained, in order to preserve the syntactic coherence of the document.

Detailed Description Text (271):
Second, the attributes of the object on which the object is being performed further dictates the cost of the operation. The attributes to be weighed into the calculation of edit distance include lexical attributes, syntactic attributes, structural attributes, and semantic attributes. Lexical attributes refer to the lexical significance of a given word; minimum content words have lesser significance and may be weighted accordingly. Syntactic attributes refer to the grammatical function of a given word, for example its part-of-speech (noun, verb, etc.). The cost of replacing one word by another may be weighted by reference to their respective syntactic attributes. Structural attributes refer to the structural significance of terms in a document (see Section IV(2)(B) above); certain known structural elements, though not insignificant, do not affect the context of the words preceding and following them. Semantic attributes refer to the semantic categorizations of certain terms, for example as company names, geographic locations or other proper nouns. These semantic categories are arranged in a hierarchy, where each such category defines a certain set of cost functions for the different editing operations performed on words of that type. For both structural and semantic attributes, domain-dependent methods for identifying these attributes are assumed. A second type of semantic attribute are synonym sets, where different terms are considered to have the same meaning. In any of these cases, the cost of a particular operation may be lowered or raised.

Detailed Description Text (311):
The smallest edit distance is 1, and either a heuristic or user interaction may be utilized to determine which of the two candidate phrases, each with identical minimal edit distance, is selected as an Approximate Phrase. Note that the particular sequence of operations associated with each minimum cost cell is recorded, to assist in determining the proper substitution, as described below.

CLAIMS:

1. A method of extracting phrases in a document, which comprise the steps of:

extracting phrases of a document to automatically create a preliminary list of extracted phrases;

filtering the preliminary list of extracted phrases to create a final list of extracted phrases;

extracting candidate phrases of the document which are similar to extracting phrases contained in the final list of extracted phrases;

confirming whether a candidate phrase of the document is sufficiently proximate to the extracted phrase to constitute an approximate phrase by calculating an edit distance of the candidate phrases based on two distinct cost functions, a first one relating to a semantic significance and role of a text of the document, and a second one elating to operations performed on the text of the document; and

computing a phrase substitution to determine the appropriate conformation of one of the extracted phrase to the approximate phrase and the approximate phrase to the extracted phrase.

2. The method of extracting phrases in a document of claim 1, wherein the step of extracting phrases of a document further comprises tokenizing the document.

3. The method of extracting phrases in a document of claim 1, wherein the step of extracting phrases of a document further comprises constructing sentence suffix trees for determination of extracted phrases.

4. The method of extracting phrases in a document of claim 3, wherein the step of extracting phrases of a document further comprises traversing and analyzing each of said suffix trees.

5. The method of extracting phrases in a document of claim 1, wherein the step of extracting phrases of a document further comprises applying a stop list.

6. The method of extracting phrases in a document of claim 1, wherein the step of filtering the preliminary list of extracted phrases further comprises extracting prefix nested phrases.

7. The method of extracting phrases in a document of claim 1, wherein the step of filtering the preliminary list of extracted phrases further comprises extracting suffix nested phrases.

8. The method of extracting phrases in a document of claim 1, wherein the step of filtering the preliminary list of extracted phrases further comprises eliminating duplicative nested phrases from the final list of induced phrases.

9. The method of extracting phrases in a document of claim 1, wherein the step of filtering the preliminary list of extracted phrases further comprises post-processing of the extracted phrase.

10. The method of extracting phrases in a document of claim 1, wherein the step of filtering the preliminary list of extracted phrases further comprises eliminating dangling words.

11. The method of extracting phrases in a document of claim 1, wherein the step of extracting candidate phrases further comprises constructing a dictionary.

12. The method of extracting phrases in a document of claim 1, wherein the step of extracting candidate phrases further comprises constructing a phrase index.

13. The method of extracting phrases in a document of claim 1, wherein the step of extracting candidate phrases further comprises extracting candidate phrases by searching for approximate extracted matches.

14. The method of extracting phrases in a document of claim 1, wherein the step of extracting candidate phrases further comprises applying a shifting window of variable starting point, ending point, and size regardless of starting words of the candidate phase.

15. The method of extracting phrases in a document of claim 1, wherein the step of extracting candidate phrases further comprises accumulating counts.

16. The method of extracting phrases in a document of claim 1, wherein the step of extracting candidate phrases further comprises generating candidates.

17. The method of extracting phrases in a document of claim 1, wherein the step of confirming whether a candidate phrase of the document is sufficiently proximate to the extracted phrase further comprises computing a phrase conforming the approximate extracted phrase to the extracted phrase.

18. The method of extracting phrases in a document of claim 1, wherein the step of confirming whether a candidate phrase of the document is sufficiently proximate to the extracted phrase further comprises computing a phrase conforming the extracted phrase to the approximate extracted phrase.

19. The method of extracting phrases in a document of claim 1, wherein the step of confirming whether a candidate phrase is sufficiently proximate to the extracted phrase further comprises weighting an edit distance.

20. The method of extracting phrases in a document of claim 1, wherein the step of extracting

phrases of a document is performed without using a pre-stored dictionary.

Previous Doc Next Doc Go to Doc#

L9: Entry 26 of 29 File: USPT Mar 20, 2001

DOCUMENT-IDENTIFIER: US 6205456 B1
TITLE: Summarization apparatus and method

Abstract Text (1):
A document summarization apparatus or method summarizes an electronic document written in a natural language, and generates an appropriate summary depending on user's focus and user's knowledge. The document summarization apparatus according to the present invention includes, for example, a focused information relevant portion extraction unit, a summary readability improvement unit, and a summary generation unit. The focused information relevant portion extraction unit extracts a portion related to two types of focused information in a document to be summarized based on the two types of focused information, that is, user-focused information as information focused by a user who uses a summary, and author-focused information as information emphasized by an author of the document to be summarized. In the document to be summarized, the summary readability improvement unit distinguishes user known information already known to a user, and information known through an access log regarded as already known to a user based on a document previously presented to the user when a summary is generated, from other information than these two types of information, and selects an important portion in the document to be summarized. The summary generation unit generates the summary of the document to be summarized based on the selection result of the summary readability improvement unit. Thus, a summary can be generated with both user-focused information and author-focused information can be included depending on the knowledge level of a user.

Brief Summary Text (9):
There have been two major methods of generating the summary of a document in the conventional document summarization technology. The first method is to recognize and extract an important portion in a document (normally the logical elements of a document such as a sentence, a paragraph, a section, etc., and hereinafter referred to as a sentence), and generate a summary. The second method is to prepare a pattern of information to be extracted as a summary and make a summary after extracting words or phrases in the document according to the condition of the pattern or extracting sentences according to the pattern. Since the second method is little related to the present invention, the first method is described below.

Brief Summary Text (14):
In method 1, that is, the method depending on the occurrence and distribution of words in a document, the importance of a word (phrase) contained in a document is normally determined first, and then the importance of the sentence is evaluated depending on the number of important words contained in the sentence. Then, an important sentence can be selected and a summary is generated. The importance of a word is calculated by using the occurrence of the word in a document, which can be weighed by taking into account the deviation of the occurrence of the word from the occurrence of the word in a common document set or the position where the word appears (a word appearing in a title is regarded as an important word, etc.). Normally, a focused word is an independent word in Japanese (especially a noun), and a content word in English. An independent word and a content word refer to a word having a substantial meaning such as a noun, adjective, verb, etc. that can be distinguished from syntactic words such as a preposition, an auxiliary, etc. The formal definition of an independent word in Japanese implies a word which can form part of an independent section in a sentence. This is a little different from the description above, but the purpose of limiting a focused word to an independent word is described above.

Brief Summary Text (18):
In the Japanese Laid-open Patent Publication (Tokkaihei) No. 07-36896 "Document Summarization Method and Apparatus", a seed for an important representation is selected based on the complexity (word length, etc.) of the representation (word, etc.) in a document, and a summary

is generated by extracting a sentence containing a larger number of important seeds.

Brief Summary Text (19):
In the Japanese Laid-open Patent Publication (Tokkaihei) No. 08-297677 "Automatic Method of Generating Summary of Subject", words of main subjects are recognized in order from the highest occurrence of a word in a document, and a summary is generated by extracting a sentence containing a larger number of important subject words.

Brief Summary Text (20):
In the Japanese Laid-open Patent Publication (Tokkaihei) No. 06-215049 "Document Summarization Apparatus", a summary is generated by extracting a sentence from a sentence or paragraph having a feature vector similar to that of the entire document after applying a vector space model often used in determining the relevance between a retrieval result and a question sentence. A vector space model refers to representing a feature of a document and a query sentence using a feature vector indicating the existence or occurrence of a word in the document and the query sentence after assigning a dimension (axis) to each keyword or each meaning element of a word.

Brief Summary Text (38):
Described below is a further problem with the linguistic document. In a linguistic document, the formal nature of a language is discussed, and the contents of an example given in the document does not have to be related to the linguistic discussion. For example, the Japanese sentences "An elephant has a long nose." is frequently cited linguistic examples. When a user searches for information about animals, a document containing such examples can be retrieved. Since the document is a linguistic document, the occurrence of words relating to animals is small when the frequency of the words in the document is checked, and it is figured out that an elephant is not an important word. If an automatically-generated summary is displayed as a retrieval result based on the frequency distribution, such examples are hardly contained in the summary, thereby makes a user confused. That is, when a keyword `elephant` is input, such a linguistic document may be retrieved, but is not contained in the display (automatically-generated summary) of a retrieval result, and the user cannot understand why such a word could be retrieved. On the other hand, when only the vicinity of a keyword is displayed, only an example is displayed and the user cannot understand what the document is about because only an example portion is displayed.

Brief Summary Text (42):
The document summarization apparatus according to the present invention comprises as components in principle a focused information relevant portion extraction unit and a summary generation unit. According to the user-focused information and the author-focused information, the focused information relevant extraction unit extracts a portion related to these two types of information from a document to be summarized. User-focused information refers to information focused by a user who uses the summary of a document to be summarized. Author-focused information refers to information which an author urges a user to pay an attention to.

Brief Summary Text (43):
The summary generation unit generates a summary of a document to be summarized based on an extraction result from the focused information relevant portion extraction unit.

Brief Summary Text (45):
The summary generation unit generates a summary of a document to be summarized using an important portion of the document to be summarized selected by the summary readability improvement unit.

Brief Summary Text (46):
The document summarization apparatus of the present invention can further be designed to include a focused information relevant portion extraction unit and a summary readability improvement unit in addition to a summary generation unit. In this case, the summary readability improvement unit selects an important portion of the document to be summarized as described above corresponding to the extraction result from the focused information relevant portion extraction unit, and the summary generation unit generates the summary of the document to be summarized based on the selection result of the summary readability improvement unit.

Drawing Description Text (18):
FIGS. 17A and 17B show an example of discourse analysis of a predicate phrase;

Drawing Description Text (24):
FIGS. 23A and 23B show an example of a summary explaining the effect of the concept knowledge criterion of a topic phrase;

Detailed Description Text (3):
FIG. 1 shows the first principle of the present invention. In FIG. 1, according to user-focused information and author-focused information, a focused information relevant portion extraction unit 1 extracts a portion related to the information from a document to be summarized. User-focused information refers to information focused by a user who uses a summary of the document. Author-focused information refers to information on which the author requests the user to focus his or her attention. A summary generation unit 2 generates a summary of the document to be summarized based on the extraction result of the focused information relevant portion extraction unit 1.

Detailed Description Text (4):
FIG. 2 shows the second principle of the present invention. In FIG. 2, according to user known information and information known through an access log, a summary readability improvement unit 3 selects an important portion in the document to be summarized using the two types of known information as being distinguished from other information to improve the readability of the summary. The user known information refers to information known to the user who uses the summary. The information known through an access log refers to information regarded as being known to the user based on the document previously presented to the user. A summary generation unit 4 generates a summary of the document to be summarized based on the selection result of the summary readability improvement unit 3.

Detailed Description Text (8):
The concept knowledge criterion indicates that an element concept composing a summary, particularly an element concept relating to a topic of a sentence, should be known. An element concept refers to a concept represented by a word appearing in a summary. In other words, a word (mainly a noun) output in a summary should be understandable for the user. Based on the criterion, the present invention supplements a plurality of parts of a document related to a word unfamiliar to the user.

Detailed Description Text (9):
The proposition knowledge criterion (or standard) relates to a combination of words appearing in a document, and indicates that as many unknown word combinations (propositions) as possible should be taken into a summary if the amount of focused information and the size of a summary is same. Based on the criterion, the present invention reduces redundant information in a summary of a document in such a manner not to select two or more sentences corresponding to a same proposition. It also reduces redundant information in a summary of a series of documents, such as a series of newspaper articles reporting a same event, in such a manner not to select two or more sentences relating to a same aspect of the event.

Detailed Description Text (15):
As for a request in generating a summary, that is, restriction information relating to generation of a summary, a required output element and other control instructions are specified by the user. A required output element refers to an element to be included in a summary. Other control instructions specify information to be used as focused information and known information, how to use the information, a desired length of summary and basic unit a summary is composed of. The basic unit can normally be a sentence or a predicate phrase.

Detailed Description Text (25):
A known concept and a known proposition are further described below. A known concept refers to a list of contents words, each word having a substantial meaning. For example, if a user knows what business Fujitsu is engaged in, the user is assigned `Fujitsu` as a known concept. If the document to be summarized begins with `Fujitsu is a Japanese computer manufacturer, and Fujitsu is planning to . . . `, the first sentence is included in a summary because the start of the second sentence `Fujitsu` implies an anaphoric expression. However, since the computer can easily determine that Fujitsu is a known concept and the first sentence only introduces the name of the company (definition of an attribute), the first sentence is not extracted according to the present embodiment.

Detailed Description Text (36):
When a plurality of documents are collectively summarized, the processes in steps S7 and S8 are performed on each document in step S14, that is, document structure information and a word list are generated. In step S15, author-focused information (author-focused words) is extracted from each document, and merged and added to the focused concept list. When the author-focused information is extracted, a document element specified as an extraction source of the author-focused information in step S1 is compared with the document structure information and a noun contained therein is selected as, for example, a focused concept. In step S16, a list of extraction units is generated in step S9, an extraction unit is selected in step S10, and a summary is generated in step S11. Then, a summary of each document is generated.

Detailed Description Text (65):
The summary formatter 15 arranges the extraction units selected by the sentence selector 14 in the order they appear in the original document, and rearranges the summary in a readable format by adding a mark indicating the existence of a sentence not extracted and by inserting a delimitation of paragraphs. If the dependence on the information known through an access log is set, a hyper-textual correlation can be set.

Detailed Description Text (68):
It is determined in step S75 whether or not an extraction unit has been retrieved. Since it is naturally determined that it has been retrieved, it is further determined in step S76 whether or not the appearance position of the retrieved extraction unit in the leading input document matches the last output position. Since the last output position is the starting position of the document in step S72, there should be a non-extracted sentence between the starting position and the retrieved leading extraction unit if it is determined that the starting position of the extraction unit does not match the last output position. Therefore, an omission element existence mark is output to the output buffer in step S77. The omission element existence mark normally corresponds to ` . . . `. However, since the mark is also used when a part of a sentence has not been extracted, it usually refers to an existence of a non-extracted element (extraction unit).

Detailed Description Text (80):
The document access log 18 accumulates the history of user's access to documents and summaries.

Detailed Description Text (84):
The discourse structure analyzing process is performed as shown in FIG. 13. The contents of a document converted by the morphological analyzer 12 into a word list are divided into extraction units to be processed by the sentence selector 14, and an extraction unit list having the divided extraction units as elements is generated. The selection restriction conditions of the extraction units to improve the readability of a summary are provided as the dependence among the extraction units according to the known concept list provided from the summarization process controller 10 and the document structure information output from the document structure analyzer 11. The restrictions to improve the readability of the summary comprise the restrictions based on the concept knowledge criterion and the restrictions based on the document structure in which a header is output corresponding to an extracted sentence.

Detailed Description Text (85):
Since the process of setting the dependence (step S96) among the extraction units in a sentence is inserted between the process of dividing a sentence into extraction units and the process of adding the extraction units to the predicate phrase list as a list of outputting the divided extraction units as shown FIG. 13, the sentence dividing process cannot be simply divided from the dependence setting process. The dividing process corresponds to the process of retrieving the leading sentence in step S92 and the process of dividing the sentence into predicate phrases encompassed by dotted lines (steps S94, S95, S97 through S100). The dependence setting process corresponds to steps S96, and steps S100 through S105.

Detailed Description Text (86):
According to the embodiment described below, a predicate phrase is used as an extraction unit in Japanese whereas a clause is used as an extraction unit in English. However, the word `predicate phrase` is hereinafter used for both Japanese and English. (Both in Japanese and English, a sentence can be processed as extraction units as described later). A predicate phrase refers to a phrase based on one predicate and a dependent noun (including a subject),

and corresponds to a simple sentence contained in a sentence. A predicate such as a word of declinable part of speech in Japanese and a verb in English is retrieved from a depending structure. Then, a dependent element excluding a predicate is added to the retrieved predicate to make a predicate phrase. A functional word, such as a conjunction, a preposition, an auxiliary, etc. is grouped with an independent word (contents word) immediately before or after the functional word. A modification element such as an adjective modifying a noun can be grouped with a modified element or can be expressed as an independent predicate phrase. However, an independent phrase should be determined depending on the wording characteristic of a modification element or the type of modification.

Detailed Description Text (87):
When the process starts as shown in FIG. 13, the contents of the predicate phrase list to be finally generated are cleared in step S91. In step S92, the leading sentence is retrieved. It is determined in step S93 whether or not the sentence has been detected. A sentence can be detected from a document using an end-of-sentence mark such as a period with the logical structure of the document taken into account by, for example, regarding a header as one sentence, etc.

Detailed Description Text (88):
If it is determined that a sentence has been detected, the structure of the syntax dependency of the sentence is analyzed in step S94. There are well-known methods of analyzing the structure of the syntactic dependency of a sentence using the dependency grammar, the phrase-structure grammar, etc. For example, the following document 10 refers to a basic method such as a chart method in which the structure of a document can be efficiently analyzed based on the context-free grammars.

Detailed Description Text (92):
Based on the analysis result of the structure of the syntactic dependency, the sentence retrieved in step S95 is analyzed into predicate phrases (simple sentences).

Detailed Description Text (93):
Whatever method is used as a method of analyzing the correspondence structure, the analysis into predicate phrases is considerably costly. However, a long sentence can be easily summarized using predicate phrases. Since a high-level meaning process is performed, a known proposition is given in a frame format shown in FIG. 15 (a frame refers to a set of a combination of an attribute name (slot name) and an attribute value (filler) and is a well-known knowledge representing method). Furthermore, a known proposition information can be compared more simply with a summary unit in the meaning network representation as shown in FIG. 16.

Detailed Description Text (95):
As described above, it is desired to selectively use a predicate phrase and a sentence as a unit of a summary., When a summary is generated in sentence units, the analysis process into predicate phrases indicated as encompassed by dotted lines in FIG. 13, that is, the processes in steps in S94 through S102, can be omitted. These processes include the analysis into predicate phrases and the settings of the dependence between the predicate phrases.

Detailed Description Text (97):
The dependence between document components is set for the sentence and phrases (subordinate sentences and phrases) which themselves have low readability but can be made more readable by taking another related sentence or phrase together into a summary. The dependence is set for the following document components.

Detailed Description Text (111):
After dividing the sentence into predicate phrases (simple sentences) in step S95 shown in FIG. 13, the dependence is set between the predicates in the syntactic dependency structure in step S96. In step S97, a predicate phrase in dependent of another predicate phrase is set as a main predicate phrase. In step S98, the main predicate phrase is added to a predicate phrase list. When a process is performed in sentence units as described above, the above described processes are omitted, and the entire sentence is regarded as a main predicate phrase. The main predicate phrase is a phrase on which another phrase depends when the dependence is set between sentences in the subsequent process. As a result, a main predicate phrase is always taken into a summary whenever one of the sentences depending on it is taken into a summary.

Detailed Description Text (112):
FIG. 17 shows an example of dividing into predicate phrases and setting the dependence. In FIG. 17A, the dependence that predicate phrase 2 in sentence 1 depends on predicate phrase 1 is set. Similarly, the dependence that predicate phrase 2 in sentence 2 depends on predicate phrase 1 is set. In either sentence, predicate phrase 1 is a main predicate phrase. As shown in FIG. 17, the dependence can be set to any pairs of predicates related syntactically each other without restriction of their relation type (i.e., direct or indirect). That is, in sentence 1, the predicate `hiita (caught)` in predicate phrase 2 is directly related to the predicate `yasunda (was absent)` in predicate phrase 1 through the conjunctive auxiliary word `node (since)`. On the other hand, in sentence 2, the predicate `okuttekureta (sent)` in predicate phrase 2 is indirectly related to the predicate `put away` in predicate phrase 1 through the noun `letter`. In these case, the dependence is similarly set.

Detailed Description Text (113):
After the main predicate phrase is added to the predicate phrase list in step S98 shown in FIG. 13, a representative phrase of a sentence is determined in steps S99 through S102. When the process is performed in sentence units, the processes in steps S99 through S102 are not required, and the entire sentence is regarded as a representative phrase (and a main predicate phrase).

Detailed Description Text (114):
A representative phrase of a sentence refers to a phrase that is an origin of the dependence set in step S104 if the sentence depends on another sentence. It is introduced mainly to treat the topic phrase (or subject phrase) of Japanese sentences.

Detailed Description Text (115):
Basic structure of a Japanese sentence is a kind of topic-comment structure. It has a topic phrase, which basically is composed of a noun followed by a topic-making postposition `wa`, and a predicate part, which is composed of a predicate, some complement phrases, and optional adverbial phrases. But in normal Japanese discourse many Japanese sentences have no topic phrase of itself, and most of them refer to the previous topic phrase of another sentence. For example, the second sentence of `Hanako-wa kaimono-ni dekaketa. Kireina huku-wo katta.` (in English, Hanako went out for shopping. .phi. bought a beautiful dress.) has no topic, and it refer to the topic phrase of the first sentence, i.e., `Hanako-wa`, as the subject of its predicate `katta (bought)`. In the terminology of the present embodiment, the second sentence depends on the first sentence.

Detailed Description Text (116):
Suppose `huku (dress)` is a focused word, both sentences should be taken into a summary because the second sentence containing the focused word needs to be taken into a summary. In this example, whole of the first sentence is not necessary for a user to recognize the second sentence. Only the topic phrase of the first sentence is enough. This is the reason why a topic phrase is separated from a predicate phrase appearing with it. The present embodiment divides the above example into three parts as follows: `Hanako-wa` (topic phrase 1: the representative phrase of the first sentence), `kaimono-ni dekaketa (went out for shopping)` (predicate phrase 1: the main predicate phrase of the first sentence), and `Kireina huku-wo katta (bought a beautiful dress)` (predicate phrase 2: the main predicate phrase and representative phrase of the second sentence), and sets the dependence of predicate phrase 1 on topic phrase 1 and of predicate phrase 2 on topic phrase 1. Thus, the second sentence does not depend on the first sentence, but depends on the topic phrase of the first sentence. In other words, the second sentence, if necessary, will be taken into a summary with its relating topic phrase just by the same mechanism to take the first sentence as a whole into a summary.

Detailed Description Text (117):
Suppose the above example be in a section entitled with `Memorandum of December 15.sup.th. Both sentences of the example depend on the title, in the present embodiment. If the focused word given is `Hanako`, the present embodiment takes topic phrase 1 and the title of the section into a summary at lease. But they are not enough because only `Memorandum of December 15.sup.th ` and `Hanako-wa` do not make a sense. They need some predicates (verb). The present embodiment avoids such a nonsense summary by giving different roles to a main predicate phrase and a representative phrase. The present embodiment takes at least one main predicate into a summary, but not a representative phrase for itself. This is a reason why a topic phrase can be a

representative phrase but not a main predicate phrase.

Detailed Description Text (118):
The present embodiment scores importance of each predicate phrase list that starts with a main predicate phrase and includes plurality of predicate phrases and topic phrases the main predicate phrase depends on, and taken plurality of predicate phrase lists into a summary. In this case, it scores following two predicate phrase lists: (1) `kaimono-ni dekaketa (went out for shopping)`, `Hanako-wa`, and `Memorandum of December 15.sup.th `; (2) `Kireina huku-wo katta (bought a beautiful dress)`, `Hanako-wa` and `Memorandum of December 15.sup.th ` A list of `Hanako-wa` and `Memorandum of December 15.sup.th ` corresponding the above described nonsense summary is no longer scored, because it does not start with a main predicate phrase. The following describes the detailed procedure of this process.

Detailed Description Text (119):
A representative phrase refers to a topic phrase or a predicate phrase which is contained in a sentence and does not depend on other predicate phrases. That is, if a topic phrase is isolated in a sentence, the topic phrase is a representative phrase. Otherwise, a main predicate phrase is a representative phrase in a sentence.

Detailed Description Text (120):
It is determined in step S99 whether or not a topic phrase exists in a sentence. If yes, the topic phrase is isolated in step S100, and the dependence is set between the topic phrase and the main predicate phrase. A topic phrase refers to a noun phrase followed by the Japanese topic maker (postposition `wa`).

Detailed Description Text (121):
FIG. 17B shows the dependence after isolating the topic phrase. In sentence 1, `Taro-wa` is a topic phrase. Predicate phrase 2 depends on predicate phrase 1, and predicate phrase 1 depends on the topic phrase. In sentence 2, `Hanako-wa` is a topic phrase. Predicate phrase 2 depends on predicate phrase 1, and predicate phrase 1 depends on the topic phrase. Thus, the analyzed predicate phrase and topic phrase are restructured based on the dependence by the sentence selector 14 described later, and when they are incorporated into a summary, they are grouped with a phrase on which they depend. By referring to sentence 1 shown in FIG. 17B as an example, what may be incorporated into a summary is `Taro-wa gakko-wo yasunda (Taro was absent from school)` (topic phrase+predicate phrase 1), or `Taro-wa kaze-wo hiitanode gakko-wo yasunda (Since Taro caught cold, he was absent from school)` (topic phrase+predicate phrase 2+predicate phrase 1).

Detailed Description Text (122):
After isolating the topic phrase and setting the dependence in step S100, the topic phrase is defined as the representative phrase of the sentence in step S101, and control is passed to the process in step S103. Unless no topic phrase exists in the sentence in step S99, the main predicate phrase is defined as the representative phrase of the sentence in step S102, and control is passed to the process in step S103.

Detailed Description Text (123):
In steps S103 and S104, the dependence between sentences set in the document structure analyzing process is converted into the relationship between predicate phrases. This process is performed only when a sentence is included in a portion subordinate to a header, etc. (dependent block of the body of a document). In step S103, it is determined whether or not the sentence being processed is an element in a dependent block. If yes, the dependence is set between the representative phrase of the sentence being processed and the main predicate phrase corresponding to the portion on which a block depend in,the document structure in step S104. Then, control is passed to the process in step S105. If it is determined in step S103 that the sentence is not an element in the dependent block, the process in step S104 is omitted and control is passed to the process in step S105. Since only typical processes are described here, there is no steps in which the dependence is set when the sentence being processed depends on a sentence after the sentence being processed. If such a process is required, a condition of specifying a dependent-on sentence and a representative phrase of a subordinate sentence should be stored and the dependence be set when a sentence satisfying the condition is processed.

Detailed Description Text (124):
In the last step shown in FIG. 13, that is, in step S105, the process of setting the dependence

and penalty is performed based on the concept knowledge criterion. According to the present embodiment, a process of setting the dependence of a topic phrase containing an unfamiliar word on the first sentence where the unfamiliar word appears, and a process of imposing a penalty on a sentence containing an anaphoric expression are performed. After this process, control is returned to the process in step S92 and the next sentence is retrieved. If it is determined in step S93 that a sentence has been retrieved, the processes in and after step S94 are repeated. If it is determined in step S93 that no sentences have been retrieved, then the discourse structure analyzing process is terminated.

Detailed Description Text (125):
The process of setting the dependence and penalty in step S105 based on the concept knowledge criterion is described further in detail by referring to FIG. 18. When the process starts as shown in the flowchart in FIG. 18, it is determined in step S111 whether or not a topic phrase exists. If yes, the predicate phrase list is searched from the beginning in step S112 for a phrase in which a main noun first appears. It is determined in step S113 whether or not such a phrase has been detected. If yes, the dependence is set between the topic phrase and the detected phrase on which the topic phrase depends, and the process terminates.

Detailed Description Text (126):
If no topic phrase is detected in step S11, or if no phrase containing a main noun of the topic phrase is detected in the predicate phrase list in step S113, then it is determined in step S115 whether or not an anaphoric expression exists at the beginning of the representative phrase. If yes, a penalty is imposed on the representative phrase in step S116. If not, the process in step S116 is not performed. Then, the process terminates.

Detailed Description Text (127):
After the discourse structure analyzing process shown in FIG. 13 has been completed, the sentence selector 14 performs the sentence selecting process. In the sentence selecting process, the sentence selector 14 selects a plurality of important predicate phrases to be taken into a summary from the predicate phrase list constructed by the discourse structure analyzer 13, and makes a list of the selected predicate phrases, called selection result list. The flowchart of the process is shown in FIG. 19.

Detailed Description Text (128):
In FIG. 19, the focused information is processed as a focused concept list. According to the present embodiment, the focused concept list comprises nouns which express important concepts (i.e., user-focused or author-focused concepts) for summarization, and the amount of focused information of a predicate phrase, which is the first measure to determine the importance of a predicate phrase, is computed by counting the occurrences of these nouns in it (describe later). Alternatively, it is possible to compute the amount of focused information by counting the occurrences of not only the nouns exactly matching an item of a focused concept list but also the synonyms of them. For example, if there is a word PC in the focused concept list, the occurrences of personal computer can be used to compute the amount of focused information. (This is a reason why the noun list is called a focused concept list, but not a focused word list.)

Detailed Description Text (130):
Then, it is determined in step S122 whether or not the essential output phrase list is blank. If not, a process of adding essential output phrases to a selection result list is performed in steps S123 and S124. An essential output phrase refers to a predicate phrase corresponding to a document element (header, etc.) instructed by the user through the summarization process controller 10 to be taken into a summary. Practically, it is a predicate phrase generated from a document element that the document structure analyzer 11 labelled as an essential output element. In step S123, the leading essential output phrase is retrieved (and removed from the essential output phrase list), and the retrieved leading phrase is added to the selection result list. In step S124, a plurality of propositions are extracted from the phrase added to the selection result list, and are added to the known proposition list, which is initially constructed by the summarization process controller, and control is returned to the process in step S122.

Detailed Description Text (131):
At this point, the focused concepts appeared in the essential output phrase can be removed from the focused concept list. However, it is better not to do so, because an essential output

element is normally a header or the like and not a complete sentence in most cases; or it is better to remove the focused concepts appeared in the essential output phrase only when the document elements from which phrase was generated is a complete sentence.

Detailed Description Text (132):
When it is determined in step S122 that the essential output phrase list is blank, a selection candidate list is generated in step S125. The selection candidate list is a list of all predicate phrases other than essential output phrases in the predicate phrase list constructed by the discourse structure analyzer 13.

Detailed Description Text (133):
In step S126, the amount of focused information of each phrase in the selection candidate list is computed. The amount of focused information of a predicate phrase is the number of focused concepts (nouns) appearing in it. If a predicate phrase has a phrase on which it depends and which has not been taken into the selection result list, the amount of focused information of the dependent predicate phrase is the sum of the number of focused concepts appearing in the both phrases (the dependent predicate phrase and the phrase on which it depends). If there are a plurality of phrases on which a predicate phrase depends, the amount of focused information of each of them is computed in advance, and the phrase having the largest amount of focused information is used to calculate the amount of focused information of the predicate phrase. When focused concepts is assigned a weight, the number is multiplied by the weight to calculate the amount of the focused information.

Detailed Description Text (134):
The amount of the focused information is computed including the phrase on which the predicate phrase depends based on the above described concept knowledge criterion. That is, according to the concept knowledge criterion, if a proper noun repeatedly appears in a document, a summary should include the first appearing portion when it includes the second appearing portion. That is, since the discourse structure analyzer 13 sets the dependence of the second sentence on the first sentence, the sentence selector 14 computes the importance of the second sentence, that is, the amount of focused information, together with the first sentence. A practical example of this process is described later.

Detailed Description Text (135):
After removing the predicate phrase of the focused information amount .phi. from the selection candidate list in step S127, the amount of new information is computed for all predicate phrases remaining in the selection candidate list in step S128. The amount of new information refers to the amount of information not to be known to a user and relating to the proposition not to be contained in the already selected predicate phrase. The computation of the amount of new information is described by referring to an example shown in FIG. 20.

Detailed Description Text (137):
Thus, in computing the amount of new information, proposition information is modelled as a concept pair (or a concept simply) and the amount of new information can be obtained by counting the propositions not contained in the already selected predicate phrase. In another method, proposition information is modelled in a format of 5W1H elements (when, where, who, what, why, and how), and is compared with predicate phrases in the frame representation shown in FIG. 15 so that the number of the predicate phrases not matching the known propositions is defined as the amount of new information. Otherwise, the amount of new information using the above described 5W1H model is referred to as the first amount of new information, and the simple amount of new information is referred to as the second amount of new information. The first and the second amounts of new information can be used in combination. In computing the amount of new information, as in computing the amount of focused information, a dependent-on predicate phrase having the largest amount of new information is selected, and the computation is made including the dependent-on phrase. For the predicate phrase on which a penalty is imposed, the amount of the penalty is subtracted from the amount of new information.

Detailed Description Text (143):
When the process in step S128 terminates, the processes in steps S131 through S136 are repeated until it is determined in step S130 that the selection candidate list becomes blank after the predicate phrase having the amount of new information of 0 is removed from the selection candidate list in step S129.

Detailed Description Text (144):
In step S131, the predicate phrase having the largest amount of focused information is selected, and the predicate phrase having the largest amount of new information is defined as an output phrase. In step S132, the output phrase is removed from the selection candidate list and added to the selection result list. At this time, if the predicate phrase has a dependent-on phrase and the dependent-on phrase has not been added to the selection result list yet, then the dependent-on phrase is added to the selection result list. If other predicate phrases having an equal amount of information exist, the predicate phrases are simultaneously added as a rule. An alternative method of selecting only one phrase based on the appearance position of a predicate phrase by, for example, selecting the predicate phrase closest to the beginning of the document to be summarized can be adopted.

Detailed Description Text (146):
Then, in step S133 shown in FIG. 19, a focused concept contained in an output phrase, that is, the predicate phrase added to the selection result list, is removed from the focused concept list. Based on the result, the amount of focused information for all predicate phrases remaining in the selection candidate list is recomputed. In step S134, the predicate phrase having the recomputed amount of focused information of 0 is removed from the selection candidate list. The recomputation of the amount of focused information can be performed as described above, and also can be performed by, for example, preliminarily storing the relationship between the focused concept and the predicate phrase, and performing the recomputation only on the predicate phrase containing the focused concept removed from the list and the predicate phrase depending on the predicate phrase added to the selection result list.

Detailed Description Text (147):
After the process in step S134, the proposition information contained in the output phrase, that is, the predicate phrase added to the selection result list in step S135, is added to the known proposition list, and the recomputation of the amount of new information is performed on all phrases remaining in the selection candidate list. The recomputation can be performed as described above, and also can be performed by, for example, storing the relationship between the proposition and the predicate phrase, and performing the recomputation only on the predicate phrase containing the proposition added to the known proposition list, the predicate phrase added to the selection result list, and the predicate phrase depending on the predicate phrase containing a changed amount of focused information.

Detailed Description Text (148):
After the predicate phrase having the amount of new information of 0 is removed from the selection candidate list in step S136 shown in FIG. 19, the processes in step S130 are repeated, and the process terminates when it is determined in step S130 that the selection candidate list becomes blank.

Detailed Description Text (149):
FIG. 21 is a flowchart showing the comparison of the amount of new information in step S131 shown in FIG. 19 when the amount of new information is divided into the first and second amounts of new information. When the amount of new information is compared between the candidate predicate phrase A and the candidate predicate phrase B as shown in FIG. 19, it is determined in step S138 which of the two predicate phrases has a larger first amount of new information. If the predicate phrase A has a larger first amount than the predicate phrase B, it is determined that the predicate phrase A has a larger amount of new information. If the predicate phrase B has a larger first amount than the predicate phrase A, it is determined that the predicate phrase B has a larger amount of new information. If the first amount of new information is equal between the predicate phrases A and B, the second amounts of new information are compared in step S139, and the predicate phrase having a larger amount of the second new information has a larger amount of new information. When the two predicate phrases A and B have an equal amount of the second new information, it is determined that these predicate phrases have an equal amount of new information.

Detailed Description Text (151):
Another factor of the length of a candidate phrase also can be used to determine which candidate phrase of those of the same new information should be included in a summary. That is, if a shorter predicate phrase is selected by priority from predicate phrases which have an equal amount of focused information and new information, a word incomprehensible to a user can be prevented from being output to a certain extent. Furthermore, instead of the comparison of

the amount of new information, the ratio (frequency of new information) of the amount of new information to the length of the selected predicate phrase can be adopted.

Detailed Description Text (156):
When a long summary is requested, an appropriate length of summary can be generated by performing a sentence selecting process according to the flowchart shown in FIG. 19 and repeating the processes in the flowchart shown in FIG. 19 on the non-selected portions. Since the phrase having the largest amount of new information is selected from the phrase having the largest amount of focused information in step S131 shown in FIG. 19, the phrase having the second largest amount of new information is selected in the second process as an output phrase. That is, based on the proposition knowledge criterion, an appropriate length of summary can be generated by taking advantage of the feature of the present invention that a redundant output is suppressed. Otherwise, a summary can be extended by sequentially fetching a more closely related portion in a method of setting all nouns in a summary obtained in the previous selecting process when the selecting process is repeatedly performed.

Detailed Description Text (167):
4. A noun contained in a selected summary portion should be used as known proposition information. That is, the number of nouns contained in the candidate predicate phrase but not contained yet in the summary is defined as the amount of new information (the number of different nouns is referred to as the first amount of new information, and the total number of the nouns is referred to as the second amount of new information).

Detailed Description Text (169):
FIG. 23 shows a practical example of generating a summary for explaining the effect of the concept knowledge criterion. It is an example of generating a summary of a economic report using the header as focused information. FIG. 23A shows a summary in which the concept knowledge criterion relating to the topic phrase `Hancock` is not used. FIG. 23B shows a summary in which the concept knowledge criterion relating to the topic phrase is used. The portion added based on the concept knowledge criterion is underlined.

Detailed Description Text (173):
3. The concept knowledge criterion is not used in FIG. 23A, but is used relating to a topic phrase in FIG. 23B.

Detailed Description Text (174):
4. A noun contained in a selected summary portion should be used as known proposition information. That is, the number of nouns contained in the candidate predicate phrase but not contained yet in the summary is defined as the amount of new information (the number of different nouns is referred to as the first amount of new information, and the total number of the nouns is referred to as the second amount of new information).

Detailed Description Text (177):
a sentence extracted also in FIG. 23A (.diamond.)

Detailed Description Text (195):
A noun appearing in the header is used as the focused information for use in generating a summary of the document to be summarized. That is, `Apple Computer`, `Windows`, `to promote`, and `to reorganize` are focused words. The amount of focused information is divided into the amount of the first focused information and the second focused information. The amount of the first focused information refers to the number of different focused words, and the amount of the second focused information refers to a total number of focused words. The amounts of the first and second focused information are processed as in the comparison of the amount of new information in FIG. 21.

Detailed Description Text (199):
Then, in step S131, sentence 11 is selected. In step S133, `reorganization` is removed from the focused word list and the focused word list becomes blank. Therefore, if the amount of the focused information is recomputed, the amounts of the focused information for predicate phrases remaining in the selection candidate list are all 0. In step S134, the contents of the selection candidate list becomes blank, thereby terminating the sentence selecting process. FIG. 23A shows the result obtained in the process.

Detailed Description Text (201):
First, when an unknown proper noun appears in a topic phrase, a dependence in which a sentence containing the proper noun first appearing in a document to be summarized is set with the sentence as a dependent-on sentence. However, in the case of a proper noun, a formal name (in this example, `Ellen Hancock` and `G. Amelio`) may be first used, but an abbreviation (in this example, `Hancock` and Amelio`) is often used from the second and subsequent occurrences. Therefore, it is regarded that a formal name equals the abbreviation. Second, when a directive word (for example, `this`) appears in a topic phrase, a dependence is set with the immediately previous sentence as a dependent-on sentence. Third, when a dependent-on sentence relates to the dependence in the first and the second processes, a dependence is set for the subsequent dependent-on sentences in the same manner.

Detailed Description Text (202):
First, the first through third processes are performed on the sentence containing a focused word, and a dependence is set. FIG. 26 shows the dependence. For example, a dependence of sentence 11 on sentence 2 is set regarding to `Hancock` as a topic phrase in sentence 11, and a dependence of sentence 2 on sentence 1 is set relating to `Amelio` in sentence 2. `Apple` and `Microsoft` in sentences 41 and 72 respectively are proper noun. However, they are famous companies and regarded as user known concepts in the following explanation.

Detailed Description Text (207):
Last, the second embodiment of the sentence selection system according to the present invention is explained. FIG. 29 shows an algorithm of extracting a sentence in this sentence selection system. This algorithm refers to the generation of digest information about articles by extracting a sentence containing a keyword of a noun using a keyword of a noun contained in the header of newspaper articles, reports, etc.

CLAIMS:

1. An apparatus for summarizing a document in support of selection, access, edition, and management of the document readable by a computer, comprising:

a focused information relevant portion extraction unit extracting a portion related to two types of focused information in a document to be summarized based on the two types of focused information comprising user-focused information as information focused by a user who uses a summary, and author-focused information as information emphasized by an author of the document to be summarized; and

a summary generation unit generating the summary of the document to be summarized based on an extraction result from said focused information relevant portion extraction unit.

9. An apparatus for summarizing a document in support of selection, access, edition, and management of the document readable by a computer, comprising:

a summary readability improvement unit improving readability of a summary by distinguishing user known information already known to a user, and/or information known through an access log regarded as already known to a user based on a document previously presented to the user when a summary is generated, from other information than these two types of information, and by selecting an important portion in a document to be summarized; and

a summary generation unit generating the summary of the document to be summarized based on a selection result from said summary readability improvement unit.

16. The document summarization apparatus according to claim 9, further comprising:

a document access log storage unit storing as a user's document access log a document and a summary presented to a user during an operation of said document summarization apparatus and a system including said document summarization apparatus, and for providing the document access log for said summary readability improvement unit as a base of the information known through an access log; and

a document cross-reference unit making cross-reference between the document and the summary stored by said document access log storage unit and the document to be summarized.

19. The document summarization apparatus according to claim 9, wherein

said summary readability improvement unit comprises:

a discourse structure analyzer for dividing each sentence in a document to be summarized into a predicate of the sentence and a predicate phrase basically including nouns depending on the predicate, defining a predicate phrase, among predicate phrases, independent of other predicate phrases as a main predicate phrase, isolating a topic phrase when the main predicate phrase contains the topic phrase, and setting a dependence between a topic phrase and a main predicate phrase and between a main predicate phrase and another predicate phrase according to a syntactic dependency structure in a sentence or between sentences.

20. An apparatus for summarizing a document in support of selection, access, edition, and management of the document readable by a computer, comprising:

a focused information relevant portion extraction unit extracting a portion related to two types of focused information in a document to be summarized based on the two types of focused information, that is; user-focused information as information focused by a user who uses a summary, and author-focused information as information emphasized by an author of the document to be summarized;

a summary readability improvement unit improving, corresponding to an extraction result from said focused information relevant portion extraction unit, readability of a summary by distinguishing user known information already known to a user, and/or information known through an access log regarded as already known to a user based on a document previously presented to the user when a summary is generated, from other information than these two types of information, and by selecting an important portion in a document to be summarized; and

a summary generation unit generating the summary of the document to be summarized based on the selection result from said summary readability improvement unit.

21. A method for summarizing a document in support of selection, access, edition, and management of the document readable by a computer, comprising:

extracting a portion related to two types of focused information in a document to be summarized based on the two types of focused information, that is, user-focused information as information focused by a user who uses a summary, and author-focused information as information emphasized by an author of the document to be summarized; and

generating the summary of the document to be summarized based on an extraction result of a portion related to the two types of focused information.

22. A method for summarizing a document in support of selection, access, edition, and management of the document readable by a computer, comprising:

distinguishing user known information already known to a user, and/or information known through an access log regarded as already known to a user based on a document previously presented to the user when a summary is generated, from other information than these two types of information, and selecting an important portion in a document to be summarized; and

generating the summary of the document to be summarized based on a selection result of the important portion.

23. A method for summarizing a document in support of selection, access, edition, and management of the document readable by a computer, comprising:

extracting a portion related to two types of focused information in a document to be summarized based on the two types of focused information, that is, user-focused information as information focused by a user who uses a summary, and author-focused information as information emphasized by an author of the document to be summarized;

distinguishing, corresponding to an extraction result, user known information already known to

a user, and/or information known through an access log regarded as already known to a user based on a document previously presented to the user when a summary is generated, from other information than these two types of information, and selecting an important portion in a document to be summarized; and

generating the summary of the document to be summarized based on a selection result of the important portion.

24. A computer-readable storage medium storing a program used to direct a computer to perform, in summarizing a document in support of selection, access, edition, and management of the document readable by a computer, the following:

extracting a portion related to two types of focused information in a document to be summarized based on the two types of focused information, that is, user-focused information as information focused by a user who uses a summary, and author-focused information as information emphasized by an author of the document to be summarized; and

generating the summary of the document to be summarized based on an extraction result of a portion related to the two types of focused information.

25. A computer-readable storage medium storing a program used to direct a computer to perform, in summarizing a document in support of selection, access, edition, and management of the document readable by a computer, the following:

distinguishing user known information already known to a user, and/or information known through an access log regarded as already known to a user based on a document previously presented to the user when a summary is generated, from other information than these two types of information, and selecting an important portion in a document to be summarized; and

generating the summary of the document to be summarized based on a selection result of the important portion.

26. A computer-readable storage medium storing aprogram used to direct a computer to perform, in summarizing a document in support of selection, access, edition, and management of the document readable by a computer, the following:

extracting a portion related to two types of focused information in a document to be summarized based on the two types of focused information, that is, user-focused information as information focused by a user who uses a summary, and author-focused information as information emphasized by an author of the document to be summarized;

distinguishing, corresponding to an extraction result, user known information already known to a user, and/or information known through an access log regarded as already known to a user based on a document previously presented to the user when a summary is generated, from other information than these two types of information, and selecting an important portion in a document to be summarized; and

generating the summary of the document to be summarized based on a selection result of the important portion.

27. An apparatus for summarizing a document in support of selection, access, edition, and management of the document readable by a computer, comprising:

a focused information relevant portion extraction unit extracting a portion related to two types of focused information in a document to be summarized based on the two types of focused information comprising user-focused information as information focused by a user who uses a summary, and author-focused information as information emphasized by an author of the document to be summarized;

a summary generation unit generating the summary of the document to be summarized based on an extraction result from said focused information relevant portion extraction unit; and

said author-focused information, refers to a title of the document, a header of a chapter, a

section, and a figure, a table of contents, and indices of words and topics, which is contained in a normally distributed document and, by which the author presents important points of the document.

28. An apparatus for summarizing a document in support of selection, access, edition, and management of the document readable by a computer, comprising:

focused information relevant portion extraction means for extracting a portion related to two types of focused information in a document to be summarized based on the two types of focused information comprising user-focused information as information focused by a user who uses a summary, and author-focused information as information emphasized by an author of the document to be summarized; and

summary generation means for generating the summary of the document to be summarized based on an extraction result from said focused information relevant portion extraction means.

29. An apparatus for summarizing a document in support of selection, access, edition, and management of the document readable by a computer, comprising:

summary readability improvement means for improving readability of a summary by distinguishing user known information already known to a user, and/or information known through an access log regarded as already known to a user based on a document previously presented to the user when a summary is generated, from other information than these two types of information, and by selecting an important portion in a document to be summarized; and

summary generation means for generating the summary of the document to be summarized based on a selection result from said summary readability improvement means.

Previous Doc Next Doc Go to Doc#